STIC Search Report
EIC 3600

TO: Michael Heck
Location: Knox 5B16
Art Unit : 3623
Thursday, August 18, 2005
Case Serial Number: 09/677993

From: Janice Burns
Location: EIC 3600
Knox / 4B71
Phone: 571-272-3518
Janice.Burns@uspto.gov

Search Notes

Dear Examiner

Here's your Fast & Focused search. Remember that it does not include all of the mandatory 705 databases, so if a full search of all databases is needed, you will have to submit the request for that separately.

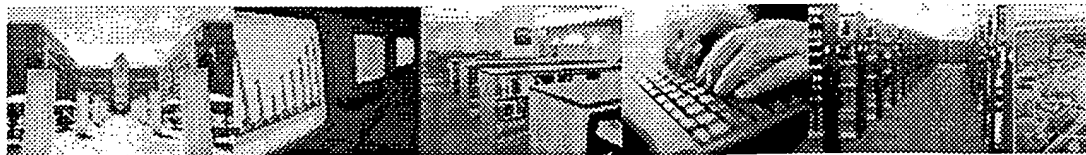If a modification or re-focus of this search is needed, please let me know.

If you have any questions please feel free to contact me.

Janice Burns, MLS
Scientific & Technical Information Center
Electronic Information Center 3600
571-272-3518
571-273-0046 (fax)
Janice.Burns@uspto.gov

BEST AVAILABLE COPY

Scientific & Technical Information Center

SERVICES

Database Search	submit
PLUS Search	submit
Book/Article Delivery	submit
Book/Journal Purchase	submit
Foreign Patents	submit
Telework Support	
Translation	submit
SIRA Automation Training	
STIC Demos & Events	

RESOURCES

STIC Online Catalog
Databases
E-Books
E-Journals
Legal Tools
Nanotechnology
Reference Tools

STIC

About Us
FAQ
Locations & Hours
News
Site Map
Staff

Search STIC Site

[ ] GO

Online Database Search Form   16 2954

Search requests relating to published applications, patent families, and litigation can be submitted by fill form and clicking on "Send."

Tech Center:

O TC 1600 O TC 1700 O TC 2100 O TC 2600 O TC 2800
O TC 2900 ◉ TC 3600 O TC 3700 O Law Lib O Other

Your Contact Information:
* indicates mandatory information.

Your Name: Michael Heck

*Email Address: michael.heck@uspto.gov
(e.g., Susan.Smith@uspto.gov)

*Art Unit/Org.: 3623

*Office Location: Knox 5B16

*Phone No.: 571-272-6730

*Case serial number: 09/677993
If not related to a patent application, please enter NA here.

Class / Subclass(es) 705/11

Earliest Priority Filing Date: 10/03/2000

Format preferred for results:
☑ Paper   ☐ Diskette   ☐ E-mail

Provide detailed information on your search topic:

- In your own words, describe in detail the concepts or subjects you want us to search.
- Include synonyms, keywords, and acronyms. Define terms that have special meanings.
- *For Chemical Structure Searches Only*
  Include the elected species or structures, keywords, synonyms, acronyms, and registry numbers
- *For Sequence Searches Only*
  Include all pertinent information (parent, child, divisional, or issued patent numbers) along with the serial number.
- *For Foreign Patent Family Searches Only*
  Include the country name and patent number.
- Provide examples or give us relevant citations, authors, etc., if known.
- FAX or send the **abstract, pertinent claims** (not all of the claims), **drawings, or chemical struc** EIC or branch library.

Enter your Search Topic Information below:

The application refers to selecting a "job posting" website that bes
fits your needs.

Specifically, I'm looking for art that ranks websites based on
infomraiton, i.e., selection criteria (preferably in a fact table or
database) using an inference engine, i.e, expert system.

I have a reference that talks about picking the right job posting
site to use, but it does not tell me how it calculates or selects th
site. (Business Wire, Webhire links corporate recruiting desktops t

Special Instructions and Other Comments:
(For fastest service, let us know the best times to contact you, in case the searcher needs further clarifi
search.)

I'm available during normal business hours

SEND   RESET

Submit comments and suggestions to <u>Kristin Vajs</u>   To report technical problems, contact <u>STIC</u>

If you cannot access a file because of a missing or non-working plugin, please contact the Help Desk at 2-9000 (Alexandria) or
305-9000 (Crystal City) for installation assistance.

<u>Intranet Home</u> | <u>Index</u> | <u>Resources</u> | <u>Contacts</u> | <u>Internet</u> | <u>Search</u> | <u>Firewall</u> | <u>Web Services</u>
Last modified 08/18/2005 10:58:01
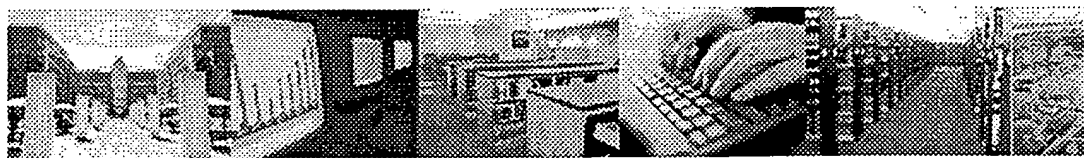
**Scientific &
Technical
Information Center**

Database Search Comfirmation

Thank you, Michael Heck. Your request (shown below) has been successfully sent to the STIC s

Your name: **Michael Heck**
Email address: **michael.heck@uspto.gov**
Art Unit: **3623**
Office Location: **Knox 5B16**
Phone_Number: **571-272-6730**

Case serial number: **09/677993**
Class / Subclass(es): **705/11**
Earliest Priority Filing Date: **10/03/2000**
Format preferred for results: **Paper**
Search Topic Information:
   The application refers to selecting a "job posting" website that best fits your needs. Specific
looking for art that ranks websites based on infomraiton, i.e., selection criteria (preferably in a f.
database) using an inference engine, i.e, expert system. I have a reference that talks about picki
job posting site to use, but it does not tell me how it calculates or selects the site. (Business Wi
links corporate recruiting desktops to over 2000 job posting sites, Business Wire, March 2, 200(
[PROQUEST])
Special Instructions and Other Comments:
   I'm available during normal business hours

Submit comments and suggestions to Kristin Vajs

To report technical problems, contact STIC

SERVICES

Database Search	submit
PLUS Search	submit
Book/Article Delivery	submit
Book/Journal Purchase	submit
Foreign Patents	submit
Telework Support	
Translation	submit
SIRA Automation Training	
STIC Demos & Events	

RESOURCES

STIC Online Catalog
Databases
E-Books
E-Journals
Legal Tools
Nanotechnology
Reference Tools

STIC

About Us
FAQ
Locations & Hours
News
Site Map
Staff

Search STIC Site

[   ] GO

If you cannot access a file because of a missing or non-working plugin, please contact the Help Desk at 2-9000 (Alexandria) or 305-9000 (Crystal City) for installation assistance.

Intranet Home | Index | Resources | Contacts | Internet | Search | Firewall | Web Services

*Last modified 08/18/2005 11:10:21*

A Web Search

Trifecta

Keeping Tabs on Search Engine Features & Technology

by Bill Mickey

3333333

a serious searcher knows that more effective results come from the same Web search query cast over several search engines. One search engine can only index so much of the Web. But as the search services become ever more competitive for end-user eyeballs, we're witness to a host of functionality refinements and licensing agreements. Database size, new relevancy ranking technologies, and a variety of query defaults and commands are just some of the issues a searcher must contend with. The Web never seems to be a known quantity, and while it's silly for a search service to purport to index all of it, all of the search engines are continually adjusting and adding to their technologies in an attempt to tame an extremely dynamic medium. Keeping tabs on these adjustments and staying informed provides a good base to construct an exhaustive search strategy.
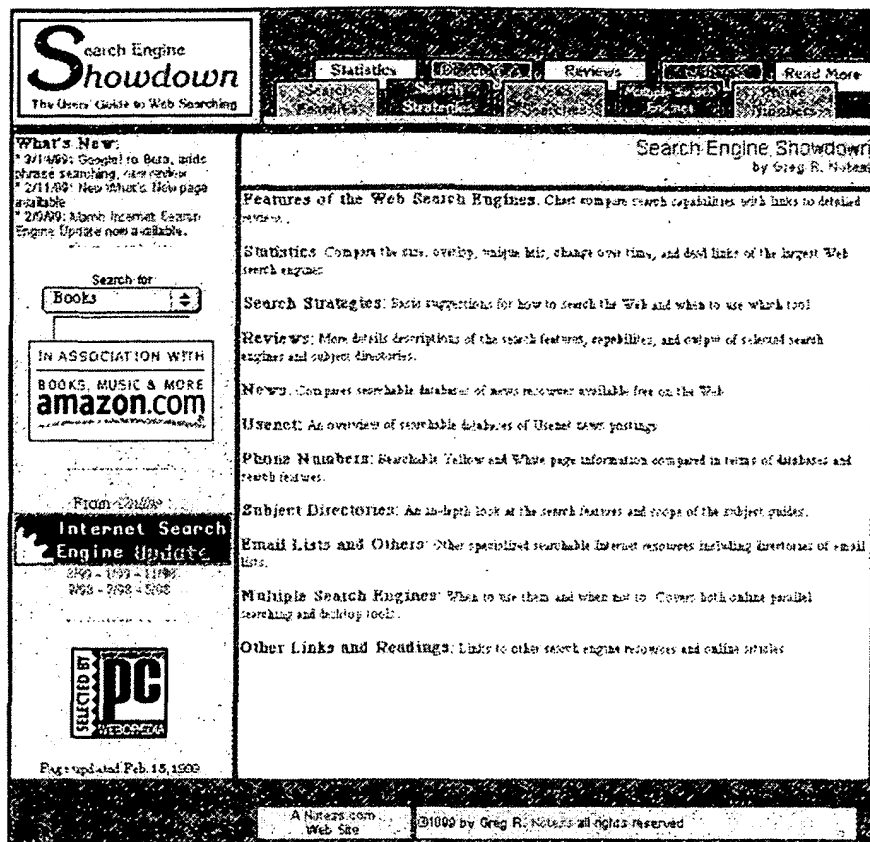
→

Search Engine Showdown
The User's Guide to Web Searching

So where on the Web do serious searchers go to keep tabs on the many new search engine developments?

Of course, one way to search several engines is to simply use a meta search engine (see Nancy Garman's article in this issue for a discussion on the current crop of meta search engines). But the bottom line is effective use of Web search engines—individually or several at a time—and effective use requires an understanding of each search engine's functionality and syntax.

So where on the Web do serious searchers go to keep tabs on the many new search engine developments? Not coincidentally, three of the authors in this issue have compiled valuable repositories of search engine information and links to other Web search-related resources. *Someone* has take the time to do statistical analyses and investigate database size, new search engine technologies, and ways to most effectively use the Web search services.

http://www.notess.com/search/

Greg Notess, ON THE NET columnist for *ONLINE* and *DATABASE* and reference librarian and associate professor at Montana State University, unveiled his new Search Engine Showdown site redesign in January. Available on the site are several useful feature tables for Web search engines, subject directories, and news search engines. Reviews pages provide further discussion of the features highlighted in the tables. Also available are feature descriptions and links to yellow and white pages, meta search engines,

and other more specialized finding aids (e.g.. for email lists, telnet resources, and gopherspace).

Notably, Greg has compiled a number of pages of statistics on the Web search services based on his own research. In fact, a January press release from Northern Light CEO David Seuss cited Search Engine Showdown's most recent study revealing Northern Light's position at the top of the database size list.

Statistics are gathered and analyzed for database size, overlap, unique hits, dead links, and database change over time. Some of the analyses were conducted at intervals over a two-year period, giving interesting insight into the evolution of the Web search engines. One of the site's highlights is the overlap page. According to Greg's test search, there was a very large percentage of unique Web pages found from the same query cast over

Notably, Greg has compiled a number of pages of statistics on the Web search services based on his own research.

...a Spam Survey reveals how well some of the major search services deal with the iniquitous practice of keyword spamming...
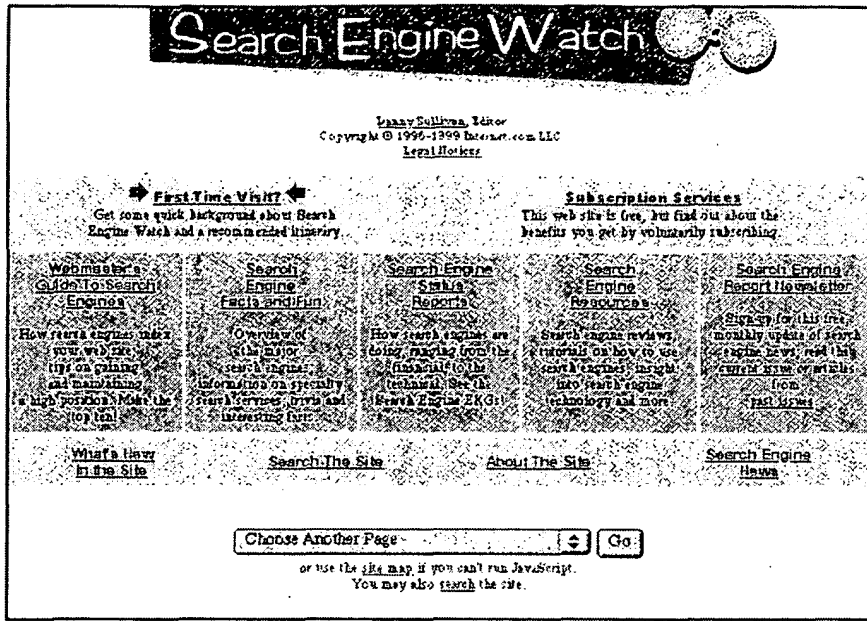
ten of the largest search engines, which included four that use Inktomi-based databases. Proof positive that at the very least, in light of Greg's findings of search engine database overlap, a multiengine search strategy is needed. Other suggestions for search strategies are noted on the Search Strategies pages.

Search Engine Watch: News, Tips and More About Search Engines

http://searchenginewatch.com/

Danny Sullivan's Search Engine Watch is part of the collection of Web sites under the Internet.com umbrella. The site's main attractions are a Webmaster's Guide to Search Engines, Search Engine Facts and Fun, Search Engine Status Reports, and Search Engine Resources. Highlighting Sullivan's commentary is a generous supply of tables, screen shots, bar charts, and graphs.

The Webmaster's Guide includes comprehensive discussion on how search engines rank Web pages and how to improve a site's ranking. This section provides an extremely well-rounded introduction to search engines and how

they work, with special attention given to Web developers.

Important sections in the Facts and Fun pages are the Guide to Search Engines and Using Search Engines. Here Sullivan provides a useful resource that compiles his own discussion of and links to many of the search engines (broken down by category, i.e., news, major, kids, regional, etc.). Additionally, Sullivan includes a collection of information on how to use search engine features, thoughtfully supported by links to tutorial resources and search engine review articles.

The Search Engine Status Reports section covers the current happenings in the search engine world from a variety of angles. For example, a Spam Survey reveals how well some of the major search services deal with the iniquitous practice of keyword spamming—important to know when performing a search that contains popular keywords that tend to be exploited by unrelated Web sites. Also included is Sullivan's clever EKG Chart, which measures the crawling habits of search engines—how often they crawl and how many pages they request from the Web sites they visit. Knowing how a search engine crawls the Web reveals much about its freshness.

In the Search Engine Resources pages, Sullivan gangs up resources for users and Webmasters. Reviews, tutorials and utilities are targeted

especially for users while Web site software, publicity resources, and design resources, among others, are presented to Webmasters. Interestingly, Sullivan includes further discussion on metatag lawsuits, highlighting current cases and their impact on Internet culture.

Web Search at Mining Co.

http://websearch.miningco.com/

As the Web Search guide for the Mining Co., Chris Sherman has put together a large collection of resource links and current commentary on hot Web search topics. In the NetLinks column on the opening page, Sherman compiled a long list of categories ranging from Arts Search to News Search to Web Site Promotion. Within each category are resource links to relevant Web sites and some even contain links to subject-specific search engines. For example, the Arts Search NetLink contains resources for education, events, galleries, news, and art-specific search engines and directories. According to Sherman, there are more than 2,500 links to specialized search resources.

The NetLinks' companion page is the Definitive Collection of Links, which provides an "ever-evolving index," complete with brief descriptions of each Net resource topic. The Definitive Collection looks to be Sherman's entire list, with 40 topics to NetLink's 30.

The Web Search site's other main attraction is In the Spotlight, which displays a number of feature articles on search-related topics. Each week Sherman adds fresh material in the form of a new feature article to the top of the list, pushing previous features down a notch. As features are bumped off the page, they're moved into the archive, which contains more than 50 articles. Occasionally, he's been known to pull a particularly popular feature out of the archives and put it back in the spotlight if current events warrant another visit to the topic. "Smart Shopping" reappeared in December and "The Investor's Web Toolkit" made a comeback in early January, for example. In mid-February, "Keeping Current with the Web" topped the list of spotlighted features. In it, Sherman provided brief commentary on links to New and Notable Sites, Informal Guides to New Resources, and Searching and Search Engine News.

LOOK SHARP

Search services tend to keep their ranking algorithms in a twilight zone of mystery and intrigue, so turn to these sites for information that you can leverage into better Web search techniques. Greg and Danny offer technical details, analysis, and news backed up by their own research of the inner-workings of Internet search engines, while Chris offers an eminently useful and comprehensive collection of topical resources coupled with his articles on new products and developments in Web search. All three repackage their news and analysis into newsletters and columns. Sullivan and Sherman offer newsletters via email and Notess reappears as the ON THE NET columnist for *ONLINE* and *DATABASE* magazines and contributes the INTERNET SEARCH ENGINE UPDATE for *ONLINE*. Both columns are available on Online Inc.'s Web site (http://www.onlineinc.com) and you can link to them from Greg's site. So between the newsletters and the resource- and research-packed Web sites, not to mention this magazine, an effective and comprehensive Web search should be well within reach.

Bill Mickey is associate editor of ONLINE *magazine. Communications to the author should be addressed to billm@onlineinc.com.*

---

miningco.com Tue, Feb 16, 1999

we mine the net so you don't have to

a guidesite for every topic

Chris Sherman - your Mining Co. Guide to:
Web Search

you are here: home > internet/online > sites/services > web search > welcome

content: welcome | netlinks | articles | guide bio | search | related
community: boards | chat | events | newsletters | feedback | share this site
shopping: bookstore | marketplace | videostore

NetLinks:
Arts Search
Business Search
Careers & Jobs
Chat Search
Clip Art Search
Computer Search
Education Search
Entertainment Search
Genealogy Search
Government Search
Health & Medicine
Image Search
Kids Search Engines
Legal/Criminal Justice
Maps & Directions
Money/Investing
MP3 Search
News Search
Online Communication
People Search
Real-Time Tracking
Reference
Religions/Beliefs
Science Search
Search Engines
Search Help & Tutorials
Search Tips & Tricks
Shopping Search
Weather Search
Web Site Promotion

More Links

In the Spotlight
Keeping Current with the Web
Stay up to date with new and notable Web sites without getting swamped by information overload.

Hot Search Sites of the Day
Great new search engines, directories, and specialty indexes—new links added every day.

Searching For Your Roots
How to search the Web for your ancestors, and get started creating your own family tree.

Web Search Myths
We debunk the most common Web search 'urban legends.'

Web Search 101
An introduction to the basics of effective Web searching, with tips on how to become a power searcher.

Super Searchers' Secrets
Great Web searching advice from four internationally recognized Masters of the Internet.

Community
Frequently Asked Questions
How to get help, how to suggest a site, and how to contact your Guide.

Web Search Newsletter
Search engine buzz, reviews, Web site promotion strategies, more. Read the Current Issue

Bulletin Board
Need search help? Ask questions or share search tips with other Web sleuths.

Web Search Chat
Open 24 hours a day. Feel free to schedule Web search chats with friends or colleagues any time.

Shopping
Bookstore
Recommendations for the best Web search, competitive intelligence, and specialized searching books.

Videostore
Browse thousands of topical videos.

Marketplace
Your source for goods and services from our sponsors.

Step 6 - Promote Your Website (Continued)
How Search Engines Rank Web Pages

So how do search engines determine your ranking? They follow a set of general rules involving the location and frequency of keywords on your Web page. These keywords may appear in your title, meta data, body text or a combination.

- Pages with keywords appearing in the title are given a higher ranking.
- Search engines also check to see if the keywords appear near the top of a Web page, such as in the headline or in the first few paragraphs of text.
- Search engines analyze how often keywords appear in relation to other words in a Web page. Those with a higher frequency are deemed more relevant than other Web pages.
- No two search engines do it exactly the same way, which is one reason why the same search on different search engines produces different results. For example:
  1. Some search engines index more Web pages than others and some index Web pages more often than others.
  2. Some search engines use link popularity as part of its ranking method, since they can tell which of the pages in its index have a lot of links pointing at them.
  3. HotBot and Infoseek give a slight boost to pages with keywords in their meta tags. But Lycos doesn't use meta data.
  4. Search engines may also penalize pages or exclude them from their index if they detect search engine spamming.

Back

Return to top

Help me improve this page, please provide me your suggestions or recommend a free new resource:
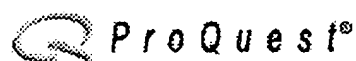
Suggestions

1st Site Free

| Home | Plan | Design | Code | Upload | Test | Promote | Maintain | Search | Map |

ProQuest®

Databases selected: Multiple databases...   What's new

THE WALL STREET JOURNAL.

Technology (A Special Report) -- Power Tools --- In Search of... Finding a better Internet search engine will go a long way toward solving the overload problem; We're getting closer

By Nick Wingfield. **Wall Street Journal.** (Eastern edition). New York, N.Y.: Jun 21, 1999. pg. R.14

Abstract (Document Summary)

Let's say, then, that a user wants to look for sites about California. The search engine goes through its database of Web sites -- millions of pages, in some cases -- and looks for sites that feature the state's name most often. Generally speaking, more mentions means a site will end up near the top of the "search results" list; fewer mentions mean a spot near the bottom.

Google also uses a robot, appropriately named Googlebot, to crawl the Web. But instead of counting how often a keyword appears on a site, Google tries to determine how highly regarded a Web page is by other Web authors.

Searching for "Star Wars" using Google, for example, should yield the site most frequently linked to by other linked-to (read: important) sites on the Web. Conversely, it should leave out, say, a Darth Vader tribute page that has won a high ranking from other search engines by wallpapering the words "Star Wars" all over the site. In this case, Google comes up with a plausible top search result: www.starwars.com, the official Web page maintained by Lucasfilm Ltd. Meanwhile, a Google search for "California politics" turns up a credible top match, the California Voter Foundation Web site.

Full Text (2307 words)

Picture the perfect Internet search engine.

It understands questions asked in conversational English -- and in Basque. And it has the brains to point you to just a handful of sites that will answer your query, not the thousand or so that usually pop up when you're looking for something online.

In short, the perfect search engine doesn't exist -- nor will it anytime soon.

There is hope, though. Search engines' frustrating shortcomings have prompted a number of companies to work on new approaches to the technology. These new engines aren't necessarily easier to communicate with, but designers think they'll help narrow down the results of searches so you won't have to sift through dozens of tangentially related sites.

The key: using people's opinions about sites instead of computers'. Some of these new search-engine schemes use surfers' own browsing habits to compile lists of the most popular and authoritative sites on any given subject. Others involve teams of researchers organizing Web compendiums by hand.

To understand how the new technologies work, you first need to understand the two current models for search engines.

First, the "robot" searches.

Engines such as Excite and AltaVista depend on "crawlers" or "spiders" -- software programs that scour Web pages and record the text they find there. The software then tallies up which words appear on which sites, and stores the information.

Let's say, then, that a user wants to look for sites about California. The search engine goes through its database of Web sites -- millions of pages, in some cases -- and looks for sites that feature the state's name most often. Generally speaking, more mentions means a site will end up near the top of the "search results" list; fewer mentions mean a spot near the bottom.

There are a number of problems with this approach. Wallpapering, for one thing. Often, to win a high ranking on a search engine, Web designers will put keywords willy-nilly all over the site.

Then there's volume. Even Net-savvy users who narrow their searches -- usually by forming a "Boolean" query, multiple words separated by and/or, such as "California and politics" -- are still flummoxed by the sheer volume of results.

"The Net is too big to be able to search effectively using the Boolean techniques developed in the '60s and '70s," says Kevin Werbach, managing editor of Release 1.0, a New York technology newsletter. In the end, Boolean queries still rely on keyword searching, so they suffer from the usual pitfalls.

Newfangled search engines look to offer something better -- by putting people's opinions, rather than keywords, in the driver's seat.

Take the new offering from Google Inc. Like Yahoo! Inc. and Excite Inc. before it, Google was founded by a group of Stanford University computer-science students who wanted to make it easier to find high-quality information on the Net. (The Palo Alto, Calif., company's name is a simplification of "googol," the word for the incredibly large number represented by a "1" followed by 100 zeroes, and is meant to convey how big the search engine's index is.)

Google also uses a robot, appropriately named Googlebot, to crawl the Web. But instead of counting how often a keyword appears on a site, Google tries to determine how highly regarded a Web page is by other Web authors.

To create its search-results list, Google counts the number of other Web pages that contain hyperlinks to that page, elevating the most-linked-to sites to the top. What's more, it looks for links only from the pool of Web pages that are themselves linked to by other sites.

But how do the sites that link to pages containing a desired keyword become important themselves? Why, naturally, by having other important sites link to them, too. If that makes your head spin, it should: Larry Page, chief executive of Google, acknowledges that the search engine's underlying logic of importance is circular, but argues it's the most effective way of finding quality information on the Net.

"What a high ranking means is that the Web as a whole believes this page is interesting or worth looking at," says Mr. Page, who says his software has analyzed more than a billion hyperlinks.

Searching for "Star Wars" using Google, for example, should yield the site most frequently linked to by other linked-to (read: important) sites on the Web. Conversely, it should leave out, say, a Darth Vader tribute page that has won a high ranking from other search engines by wallpapering the words "Star Wars" all over the site. In this case, Google comes up with a plausible top search result: www.starwars.com, the official Web page maintained by Lucasfilm Ltd. Meanwhile, a Google search for "California politics" turns up a credible top match, the California Voter Foundation Web site.

International Business Machines Corp. has its own answer to Google: Clever, a search engine under development at its Almaden Research Center near San Jose, Calif. Clever also uses hyperlink analysis to score search results, but IBM says it's better than Google at digging out high-quality sites relegated to obscurity because other high-quality sites don't link to them.

Prabhakar Raghavan, manager of computer-sciences principles at the Almaden Research Center, says the Web is littered with countless examples of such sites. "Semiconductor-manufacturer pages aren't linked together, but there is a graduate student in Calgary who's written a fantastic history of semiconductors and points to all these pages -- yet no one points to him," Mr. Raghavan says. "He wouldn't get a high score on Google."

What Clever does is trace the hyperlinks pointing to a set of Web sites back to their point of origin, so that it can rank the latter page higher. So, if you were doing a general search for semiconductor information, the student's page would show up more prominently than the manufacturers'.

Google's Mr. Page, for his part, argues that an obscure site with top-notch content will eventually get discovered by the Web at large. When that happens, he says, Google will follow the new hyperlinks pointing to the site.

For now, we'll have to trust IBM's description of Clever: Only Big Blue employees can currently access the search engine, which hums away on three dinky PCs in a room at Almaden. IBM says it's talking to a number of Web portals about licensing Clever.

While Google and Clever make Web-page authors the key players for determining the rankings, Direct Hit Technologies Inc., of Wellesley, Mass., puts the scoreboard in the hands of the Web-surfing masses -- but without making users do any work. Direct Hit piggybacks on traditional keyword search engines, kicking in after a user gets a list of search results. By monitoring the amount of time users spend on the sites yielded by their list of search results, Direct Hit comes up with a rough gauge of the sites' popularity, giving the sites that are visited longer higher rankings in future searches.

Direct Hit is getting more personal, too: The company is working on a version that will bug readers for bits of demographic data, such as gender and age. That information, the company says, is useful in ranking searches for different audiences. For instance, the company's research shows men and women are usually looking for different things when they enter "flowers" as a search term: Women generally want information about seeds and garden gadgets, while odds are men want information about where to buy bouquets.

Direct Hit isn't exactly a window onto users' souls, though. Let's say you do a search on HotBot for "California wine" and then click off to the California Wine Club Home Page, the first item on the search list. If you don't come back to HotBot, Direct Hit assumes you're happy with the California Wine Club, and that site gets bumped up in the rankings. But if you don't find what you're looking for and quickly click back to HotBot in disgust, that action penalizes the California Wine Club in future searches on the topic.

In other words, Direct Hit infers a lot from a little vague behavioral data: Maybe the surfer loved a site, or maybe she just went to the bathroom after making the query. Company co-founder Gary Culliss, though, says such situations are "statistical aberrations" that don't skew search results.

Tracking a user's every move once she leaves the search engine might yield great insights into the relationships between sites, but Mr. Culliss says it's not an option because of user-privacy concerns.

Still, Direct Hit is an effective way of harnessing people power for better search results, says Mr. Culliss, a 28-year-old former patent attorney who likes to call Direct Hit a "popularity engine."

"Everybody realizes that by adding human intelligence to search results, you get better answers," says Mr. Culliss.

But other experts expect people to play an even more direct role in creating the search engines of tomorrow -- as editors. Which brings us to the second major model for Web searching: Yahoo.

Technically, Yahoo isn't a search engine at all; it's a Web directory. Yahoo does its site-finding not by robot but by "taxonomists" -- surfers who peruse the Web and organize the sites they find into lengthy lists of categories and subcategories. What you end up with is a smaller, but more focused, list of sites -- since, obviously, a person can make better distinctions than a piece of software can.

There's a problem with Yahoo's system, though: No matter how many eager young taxonomists Yahoo gives stock options to, it can't keep up with the quantity of pages on the Web.

That's what keeps Mr. Culliss, for example, from using editors at Direct Hit. Hiring a big staff of surfers is "really expensive and not really scalable," he says. "Ultimately, it makes sense for users" to organize search results.

But some experts believe hand-picked directories will always be more satisfying for consumers. "Everybody but Yahoo started out saying, `We think machines can do it all,'" says Danny Sullivan, editor of the Web site Search Engine Watch. "I don't think machines can cut it for the bulk of general-purpose searches."

If Yahoo can't crack the "scalability" nut -- as tech-heads call the challenge of handling loads of information -- a movement called Open Directory may get a little closer to the goal. Originally developed by NewHoo!, a company later acquired by America Online Inc.'s Netscape Communications division, Open Directory is something like taxonomy on growth hormones: It relies on more than 10,000 volunteer editors, rather than paid staffers, to classify sites in tens of thousands of categories. While Yahoo doesn't disclose how many editors it has, the company as a whole has fewer than 1,000 employees.

Some experts see NewHoo! as the start of a new trend in directories: piling on personnel instead of technology. Some observers even see a dark side to the effort.

The need for in-house editorial teams to keep up with the Web's growth will make Web companies "the sweatshops of the future," says Andrew de Vries, director of marketing communications at HotBot.

With about a half-million sites, Open Directory is gaining fast on Yahoo, whose directory includes more than a million sites.

Yahoo, though, argues size doesn't matter as much when it comes to directories -- quality does. And a coordinated team of in-house editors can produce better results than a fleet of volunteers, according to the company.

"Our goal has never been to get every word on every Web site online," says Srinija Srinivasan, Yahoo's editor in chief.

Even traditional search engines admit they are going to need help from the new people-powered searching techniques. In fact, they're already getting it: A number of mainstream search engines, including Excite, are using Google-style link analysis in combination with keyword searching to improve searches for users. Lycos Inc., whose name refers to the mechanical "spider" it created to crawl the Web, has begun using Open Directory, and Yahoo has sought to back up its thinner listings with a beefier search engine from Inktomi Corp. HotBot, which is being acquired by Lycos, uses Direct Hit to provide its top-10 results in most cases.

But those efforts still seem light-years away from the ultimate science-fiction search engine: the computerized librarian that receives spoken commands and questions, then fetches flawless answers. In our fantasies, the question, "Who is on the board of directors at Microsoft?" would be quickly answered with a list of the correct names.

There are some early attempts to create a dialogue between the searcher and search engine: Ask Jeeves Inc., a Berkeley, Calif., company, operates a public Web site that invites users to type questions into its search engine. Ask Jeeves tries to match the questions to a database of about seven million similar-sounding questions to which it already knows the answers. That might seem like a lot, but it's not hard to stump Ask Jeeves: For instance, it doesn't even come close to accurately answering the question about Microsoft directors, though it did offer hints on where to find Microsoft stock quotes and information about the company's software.

How good can Jeeves get? Improvements in its answering capabilities are mostly dependent on the work of human editorial teams, who add new questions. No matter how big the team, anticipating every possible question out there is an almost unimaginable task.

Information-retrieval experts are optimistic that smarter engines are on the way. Lexeme Inc., of Cambridge, Mass., is one of a number of companies working on techniques for automatically parsing questions, without the help of human editors. Such a process requires massive computing horsepower, but Lexeme Chief Executive John Clippinger says improvements in computer microprocessors are slowly nudging search engines along towards better intelligence.

"This is not a trivial task," he says.

Mr. Wingfield is a staff reporter for The Wall Street Journal Interactive Edition in San Francisco.
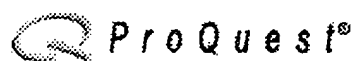
ProQuest   « Back to Document View

San Francisco Chronicle

Unraveling the Web / New search engines scan more sites, rank popularity and use human expertise for better results; [FINAL Edition]

Deborah Solomon, Chronicle Staff Writer. **San Francisco Chronicle**. San Francisco, Calif.: Aug 30, 1999. pg. C.1

Subjects:	Web sites, Search engines, Internet, Trends
Author(s):	Deborah Solomon, Chronicle Staff Writer
Document types:	News
Section:	*BUSINESS*
Publication title:	San Francisco Chronicle. San Francisco, Calif.: Aug 30, 1999. pg. C.1
Source type:	Newspaper
ProQuest document ID:	44315467
Text Word Count	2210
Document URL:	http://proquest.umi.com/pqdweb?did=44315467&sid=-1&Fmt=3&cli entId=19649&RQT=309&VName=PQD

Abstract (Document Summary)

Some are searching ever-larger portions of the Web. Others are employing staffs of editors to hunt down the best sites for particular queries. A few are even running "popularity" contests, letting Web surfers and Web-page designers guide one another to the best sites.

Traditional search engines such as Excite and AltaVista use software programs that search for Web sites containing whatever keywords the user has entered into the search bar.

-- Google of Palo Alto is a search engine that ranks Web sites based on how popular they are with other Web authors. For example, the more sites that include links to a particular page like Joe's Home Page, the more likely it is that Joe's Home Page will pop to the top of Google's search results.

Full Text (2210 words)

Finding information on the Web has always been frustrating, and as the amount of data on the Web explodes, the search is only going to get harder.

But a new breed of search engines is aiming to ease the aggravation.

Some are searching ever-larger portions of the Web. Others are employing staffs of editors to hunt down the best sites for particular queries. A few are even running "popularity" contests, letting Web surfers and Web-page designers guide one another to the best sites.

Always eager for the next best thing, Internet users are heading to the new sites in droves, giving veteran search engines a run for their money. LookSmart, Ask Jeeves and GoTo -- all fairly new sites - - ranked among the top 10 search sites for July 1999, according to Media Metrix.

To use the new technologies, it helps to understand the old ones.

Traditional search engines such as Excite and AltaVista use software programs that search for Web sites containing whatever

keywords the user has entered into the search bar.

These programs -- known as "spiders" or "crawlers" -- visit a Web page, read it and record the words on each page. The spider then makes a list of which words appear on which pages and returns those pages whenever a user types in that keyword. Generally, the more times a keyword appears on a page, the higher it ranks on a list of results.

But search engines don't always produce the best results.

For one thing, many Web site designers "wallpaper" their pages, loading them up with keywords so they'll jump to the top of a search- results list.

Also, unless you're a skilled searcher, you'll get thousands of irrelevant results. For example, if you type the word "weddings" into a search engine, you're likely to get photos of people's weddings, wedding photographers in Iowa and men looking for wives.

Another type of search site -- called a directory -- separates search results into user-friendly categories.

For example, type "weddings" into Yahoo, a direc- tory, and you'll get choices such as wedding rings, gowns and even movies (including "The Wedding Singer").

Some of the new sites are search engines and some are directories, but all claim to go a step beyond the existing technology.

-- AlltheWeb, which launched in May, aims to access more of the Web than any other search engine. It was the first to break the 200 million Web-page barrier and claims that it will access "all the Web" -- 800 million pages -- in six months.

-- Google of Palo Alto is a search engine that ranks Web sites based on how popular they are with other Web authors. For example, the more sites that include links to a particular page like Joe's Home Page, the more likely it is that Joe's Home Page will pop to the top of Google's search results.

Google's founders say it's a democratic process that lets the Web community determine which pages are worthy.

It was created by two Stanford University students, Sergey Vrin and Larry Page, who were frustrated by the existing choice of search engines.

"There are a number of other companies like Excite and Infoseek and they have search components, but primarily, they are media companies," said Vrin. Google is a no-frills search engine that aims to do one thing well.

-- Direct Hit also uses a "popularity engine" to deliver Web sites. When users key in a search, Direct Hit anonymously monitors which sites they access and how much time they spend there. The more often a site is accessed and the longer it is used, the higher its ranking.

-- Ask Jeeves of Berkeley lets users pose questions in plain English. Ask Jeeves then directs users to sites that provide the best answers. (For a complete explanation, go to www.askjeeves.com, click on "Popular Questions" and then "What Is Ask Jeeves.")

Ask Jeeves does a good job answering simple, common questions, such as "What is the capital of New Hampshire?" or "How high is the Empire State Building?" But often it's stumped by more difficult queries.

-- LookSmart of San Francisco is a directory that has more than 200 editors who scour the Web, search out the best sites and add them to a growing directory of more than 800,000 unique pages in 60,000 categories.

"The way we differentiate ourselves is in the actual size and quality of our site," said Val Landi, senior vice president of marketing and media. "We believe we have the largest staff of editors, and the assumption is that, as opposed to our robotic brethren out there, if you have intelligent, professional editors, they can select the best quality Web sites."

While there are no scientific studies that prove which type of search engine produces better results, industry watchers say the Googles and LookSmarts of the world do a good job.

The competition from these upstarts has not gone unnoticed. In recent weeks, some of the original search engines have announced planned improvements.

Excite announced plans to access more Web pages -- up to 43 percent of the Web, compared with about 6 percent now -- and is pairing with LookSmart to provide a directory.

Netscape announced that it would use Google's technology to power its search function.

AltaVista's new owner, CMGI Inc., vowed to improve the search function and turn the site into a "megaportal" where users can also do things like buy a car and trade stocks.

Microsoft also plans to announce a major revamping of its MSN.com search site in early fall.

"We are developing a next-generation search engine that leverages a lot of our experience in making software easier to use to deliver a search experience that is everything you never thought search could be," said Rob Bennett, director of marketing for MSN.

Industry watchers say members of the old guard need to improve their search functions if they want to keep users coming back.

"There is no allegiance out there," said John Corcoran, an analyst with Stephens Inc. in Boston. "Consumers are saying, `If I can get the information with only three clicks on Yahoo, fantastic. But if I can get it with two clicks on a newer search engine, I'll use that one.' "

Getting repeat customers is crucial if Net companies want to attract advertising and e-commerce partners.

"Search is a business. It clearly generates revenue because it's one of the best opportunities to reach people who are looking for something in particular," said Barry Parr, who tracks search engines for International Data Corp.

While some of the older search sites have transformed themselves into portals, some of the newer ones, including AlltheWeb and Google, plan to make money by licensing their technology to other search engines and to companies that want a search tool on their Web sites.

"Our focus is narrow: We want to do core search technology better than anyone else in the world, then take that technology and sell it to big companies," said David Burns, president and CEO of Fast Search & Transfer, which operates AlltheWeb.com.

With all the choices out there, Web searchers may be more confused than ever. Here's a piece of basic advice:

"If you don't know where to begin, directories are good places to start," said Danny Sullivan, editor of SearchEngineWatch.com in London.

When the hunt is for something very specific or obscure, like information on a rare medical condition, then a search engine may be the better choice.

"A search engine is good because you're getting into the nooks and crannies of the Web," Sullivan said.

A recent study published in the journal Nature prompted debate when it revealed that most search sites access just a tiny fraction of the Web. Northern Light accessed the most: 16 percent of the Web, compared with just 5.6 percent for Excite. AlltheWeb searches 25 percent but was not included in the study.

But unless you're looking for something rare, chances are you don't need to access much more than 15 percent of the Web.

"Most people don't say, `I wish I had another 1 million pages to look through,' " Sullivan said.

Parr of IDC agreed. "If you're doing a general search and you get 1 million pages back instead of 150,000, you're only going to look at the first 20 anyway."

Eventually, the number of search sites will stop growing and begin to shrink.

"The market will have a shakeout," Corcoran said. "Why do we need 45 search engines? We don't. The big ones now will continue to get bigger and will continue to spend large amounts of money growing their subscriber base. And the end result will be a couple of really big, really good sites."

TOP WEB SEARCH SITES

There are different ways to search the Web. Search engines, like AltaVista, crawl the web and record the text on every Web page. When you make a query, the search engine goes into the depths of the page to find relevant keywords.

A directory, such as Yahoo, is an organized selection of categories, such as Travel and Food. The content within those categories has been hand-picked by humans, who scour the Web, looking for the best sites. When you submit a query, it pulls up relevant sites just from the ones that are included in the directory.

Name / Address / Type / How much of the Web it searches / Comments

AltaVista / www.altavista.com / search engine - 15.5% / One of the oldest search engines, recently invested in by CMGI.

Excite / www.excite.com / portal / 5.6% / Plans to add more pages to its search and access about 50% of the Web.

AllTheWeb / www.alltheweb.com / search engine / 25% / Plans to increase to 100% over next year.

Northern Light / www.northernlight.com / search engine - 16% / Named for a clipper ship built in Boston in 1853 and known for its technology.

Google / www.google.com / search engine / 7.8% / Ranks Web sites bases on how often they're linked to from other sites.

LookSmart / www.looksmart.com / directory / N/A / Has more than 60,000 categories of information.

Yahoo / www.yahoo.com / directory / 7.4% / Most visited search site -- more than 38.9 million visitors.

Lycos / www.lycos.com / directory / 2.5% / Recently switched from a search engine to a directory model.

Go Network / www.go.com / portal / 8% / The Mickey Mouse Corp. soon will own all of go.com.

Ask Jeeves / www.askjeeves.com / search engine / 7 million answers / lets users make queries in form of a question.

Chronicle Graphic

SIMPLE SEARCH STRATEGIES

Most, but not all, search engines incorporate Boolean searching as an advanced form of linking keywords. The results will narrow the number of pages pertaining to your keywords. Sometimes, if you look carefully on the search page, you can find an icon or phrase that will say "advanced searches." This will lead you to a page that will accept full Boolean terms. Also, there usually are instructions included on that Web page.

-- When making a search, analyze your idea and try to state it clearly in a simple sentence.

SEARCH TERMS / WHAT YOU ENTER IN SEARCH BOX

I want to buy a dog, but I'm not sure of what breed.

AND or (+) / dogs AND buy AND puppies AND breed

Finds Web pages containing all keywords +dogs +buy +puppies +breeds

If I buy a dog, how will I groom and take care of it?

OR / buy OR purchase AND dogs OR puppies AND grooming OR cleaning AND care OR raising

Finds Web pages containing any and all keywords. Using synomyns will increase your chances of finding pages

"When taking care of your new dog"

"Phrase search" / "when taking care of your new dog"

Use quotation marks to look for a specific phrase

I want to buy a dog, but I'm not sure what breed.

AND NOT or (-) / dogs AND buy AND breed AND NOT "School of" AND NOT "School" +dogs +breeds +grooming -"School of" -"School"

Retrieves Web pages containing one keyword but not the other. Helps in reducing the number of pages youy don't want. Make sure to use quotation marks if keywords could be specific phrase.

.

What kind of dog breed

Asterik (*) / dog and breed* (will result in breed, breeds, breeding, and breeders of dogs)

By truncating a keyword and placing an asterik at its end, search engines find variations for the word.

.

I want Beagles, Dalmations and Bulldogs

Comma (,) / Beagles, Dalmations, Bulldogs

A comma is used to separate and search for proper nouns

:

I want to find a Web page "Raising and caring for your dog"

Title search (title:) / title: Raising and caring for your dog

Searches the titles of Web pages. .

-- Remember to spell keywords properly. Sometimes the Web will have different ways of spelling the word.

-- Search engines recognize keywords in all lowercase letters as either uppercase or lowercase letters. If you use initial capital letters or all capital letters, the search engine will return only pages that match your keyword exactly.

.

Sources: U.C. Berkeley, Chronicle research

Chronicle Graphic

[Illustration]
GRAPHIC; Caption: GRAPHIC: DAN HUBIG/The Chronicle

ProQuest®

Databases selected: Multiple databases...

What's new

THE WALL STREET JOURNAL.

The Best Way to.... ....Search Online: Finding what you need on the Web is getting easier and easier; (But it's still not easy)

By Timothy Hanrahan. **Wall Street Journal.** (Eastern edition). New York, N.Y.: Dec 6, 1999. pg. R.25

Subjects:	Series & special reports, Search engines, Web sites
Author(s):	By Timothy Hanrahan
Document types:	Feature
Section:	*The Internet (A Special Report)*
Publication title:	Wall Street Journal. (Eastern edition). New York, N.Y.: Dec 6, 1999. pg. R.25
Source type:	Newspaper
ISSN/ISBN:	00999660
ProQuest document ID:	46877734
Text Word Count	2208
Document URL:	http://proquest.umi.com/pqdweb?did=46877734&Fmt=3&clientId=1 9649&RQT=309&VName=PQD

Abstract (Document Summary)

And while there's a crowded field of companies engaged in cataloging the millions of Web pages popping up daily, Jupiter analyst David Card says that "there's still a way to differentiate yourself via search and directory services to form a best-of-breed search offering." Some of these new companies are trying to do that by adding a human element to the search process. Many others use conventional "spiders" -- software tools that prowl the Internet in search of new sites -- but add some technological twist.

SavvySearch.com, for instance, is a "metasearch" engine, combining and ranking results from more than 200 sources -- not just other search engines, including the big ones, but also auctions and newsgroups. (SavvySearch was recently acquired by CNET Inc., which wants to use it to bolster its computer-related search functions.) What's different about Google Inc.'s Google.com search engine is how it ranks the sites that a particular search turns up: The ranking is based on how many other sites have links to them. A ranking system is also what sets apart Direct Hit Technologies Inc.'s engine. It ranks sites according to how much time previous searchers spent at them.

Search engines don't actually search the Web instantaneously upon request; rather, they search the pages its spiders, or people, have already brought back. (An electronic spider can take days to do a complete run of the Internet, and people, of course, are even slower.) That's one reason why there is so much difference between the results of two different search engines, and why the results may change from search to search with the same engine.

Full Text (2208 words)

Still having trouble finding what you're looking for online? Several of the Web's second-tier search sites hope so.

The Web's first wave of search engines and Web directories -- Yahoo!, Lycos, Excite and others -- pushed far beyond search years ago, adding everything from free e-mail to games and chat. Their goal: to turn themselves into "portals," all-purpose home bases for Web users.

But several smaller companies have decided there's money to be made in specialization. Searching, after all, remains extremely popular -- an August survey by New York market-research firm Jupiter Communications found that 88% of the online population uses search engines; only e-mail is more widely used.

Given that, these companies see a demand for search engines that can save users time by yielding quicker and more useful results. They believe they have found better ways of pointing users to sites that actually hold the information they seek, rather than

to a hodgepodge of sites on vaguely related topics -- an all-too-common problem with search engines.

And while there's a crowded field of companies engaged in cataloging the millions of Web pages popping up daily, Jupiter analyst David Card says that "there's still a way to differentiate yourself via search and directory services to form a best-of-breed search offering." Some of these new companies are trying to do that by adding a human element to the search process. Many others use conventional "spiders" -- software tools that prowl the Internet in search of new sites -- but add some technological twist.

SavvySearch.com, for instance, is a "metasearch" engine, combining and ranking results from more than 200 sources -- not just other search engines, including the big ones, but also auctions and newsgroups. (SavvySearch was recently acquired by CNET Inc., which wants to use it to bolster its computer-related search functions.) What's different about Google Inc.'s Google.com search engine is how it ranks the sites that a particular search turns up: The ranking is based on how many other sites have links to them. A ranking system is also what sets apart Direct Hit Technologies Inc.'s engine. It ranks sites according to how much time previous searchers spent at them.

Another site, GoTo.com Inc., relies on a less sophisticated method: Web sites bid to get the top spot in its search results.

About.com Inc.'s site is one of those with a human touch: It has 650 so-called guides -- they even get head shots on the site -- who keep track of what the Web offers on 20,000 topics. They track their corner of cyberspace, writing short overviews and constantly revising links to relevant sites.

Search engines don't actually search the Web instantaneously upon request; rather, they search the pages its spiders, or people, have already brought back. (An electronic spider can take days to do a complete run of the Internet, and people, of course, are even slower.) That's one reason why there is so much difference between the results of two different search engines, and why the results may change from search to search with the same engine.

People are slower than computers, but they provide context, About.com says. Indeed, the guides go beyond pointing to appropriate sites; if they feel it's needed, they write their own material.

"Our guides look for what's out there, and if it isn't good, they create it," says Scott Kurnit, About.com's chairman and chief executive. "That's the advantage."

Competition among the new search offerings, and the innovations that has spawned, has led to a "sort of a renaissance in search," says Mike Cassidy, co-founder and CEO of Direct Hit.

We put that assertion to the test, running the same searches on six different sites: five of the new breed plus familiar old AltaVista, now a unit of CMGI Inc. (AltaVista, which gets 1,000 queries a second, recently upgraded its search capabilities to include relevancy rankings.)

The aim of a search engine is to return a useful link, and to have it near the top of the list, where a Web user is more likely to see it. With that in mind, we considered the first 10 links each engine brought up for each test; in most cases that was the first full page -- though SavvySearch provides 15 on a page, Google allows the user to choose 10, 30 or 100 and GoTo.com comes up with an ungainly 40 per page. Our questions: How many of the top 10 choices are suitable to our purposes? How suitable are they?

Search 1

Riding the PATH train

In search of schedules for the PATH trains connecting New Jersey and New York City, we tried typing "path train" on each of the six engines. That's not as straightforward a challenge as it might seem. Both "path" and "train" are relatively common terms (PATH in this case is short for Port Authority Trans-Hudson), so finding sites that relate to the trains we were interested in is a test of a search engine's discernment. And given the likelihood that not many sites are devoted to the PATH system -- compared with, say, the number devoted to the New York City subway system -- it's also a test of the engine's searching capacity.

As it turns out, all six search engines were tripped up to some degree -- but each managed to come up with at least one site in its top 10 that was within a click or two of the PATH train schedule. There seem to be two versions of that -- identical information, different layouts -- one on the Port Authority's own site and one on a site called New Jersey Online.

Direct Hit came off as the star: Six of its first 10 links led in short order to one of the schedules, and that's not taking into account that atop its list of links is a list of suggested terms for refining the search, including "Path Train Schedules" (though after clicking on that, it actually took two more clicks to get to the schedule itself).

Google claimed to have found 88,245 links, topped by one that was two clicks away from New Jersey Online's PATH schedule. Google's No. 2 choice would actually get you there one step faster, and the No. 4 choice was just one click away from the Port Authority version.

About.com's first three links all led to a PATH schedule, though the design of the site had the effect of making the schedule one click farther away than it was with the other search engines. None of About.com's other seven links were PATH-related.

GoTo.com's top result led to the Port Authority's PATH schedule and No. 2 led to New Jersey Online's, but only one of the other links opened the way to a schedule. AltaVista was just two for 10 -- one link to each of the schedules -- and those two were a bit buried, down in the seventh and ninth positions.

And then there's SavvySearch. At the head of its page of links is a list of the search engines it used for the metasearch; in this case, it used 10, including two of the engines we were comparing it with -- AltaVista and Direct Hit. Even so, just one of its top 10 links would be of any use to a PATH-schedule seeker. Fortunately, it was No. 1 on the list.

Most of the six engines' misses were easy to understand. AltaVista, which has a reputation for casting a wide net, was most prone to point to sites that mention the PATH system in passing; a number of businesses and institutions, for example, provide directions that refer to the PATH train. AltaVista was also caught by a site promoting a recording called "PATH Train." SavvySearch and Direct Hit were both tripped up by a site for Career Path Training Corp., which contains the magic phrase "path train."

Google turned up a site devoted to photos of the PATH system -- but also seemed drawn to sites that merely involved trains, including one containing a story about a commuter train crash in London.

GoTo.com was most prone to pointing to sites that contained the words "path" and "train," though not necessarily right beside one another (a site demonstrating "the easy path to horse training," for example). Putting "path train" in quotation marks eliminated that problem, but didn't increase the number of suitable sites in the top 10; in fact, it cut it to two from three. All that changed was the nature of the misses: They were now sites that mentioned the PATH train only in passing. Quotation marks weren't used for searches on any of the other engines -- though AltaVista recommends it, in this case it didn't make any difference -- but on SavvySearch we did press the "phrase" button. (In the default setting the "and" button is on.)

As for About.com, having three useful sites in the top 10 was a respectable showing, and having them in the first three spots made it even more so. But there was something alarmingly random about the other seven sites: two on mountain biking in greater St. Louis (a combination of "bike path" and a reference to training explains that), one on unconventional honeymoons (there's something about train travel, but that's pretty weak) and four along the lines of "Diary of an Exhibitionist in Italy" (go figure).

Search 2

Understanding Glass-Steagall

Next we looked for information on the recently passed repeal of the complex Glass-Steagall Act, a step that will deregulate the banking industry, by typing in "glass-steagall." Unlike the PATH search, in which we knew exactly what we wanted, here we were interested in what the engines might dig up.

This turned out to be a case in which About.com's human intervention came in handy. All of its 10 links were to articles by the search engine's banking guide, Kathy Durham -- described as a banking professional for 20 years, most recently a vice president with Bank of America in Hawaii -- and many were following the progress of banking-reform efforts. Though the articles were listed oddly, with the latest article on the passage of the bill down at No. 9, it was still relatively easy to follow the story. At the bottom of that were further links to articles by Reuters, ABC, CBS and the Los Angeles Times.

Beside that, the machine-made efforts looked a bit disorderly -- and largely indistinguishable from one another. AltaVista, GoTo.com and SavvySearch, for example, all had links to a few paragraphs on Glass-Steagall at the J.P. Morgan & Co. Web site, making a case for repeal. Those paragraphs don't mention the bill Congress just voted on, though; instead, they refer to Congress's repeal attempts of 1988 and 1991. SavvySearch, Google and Direct Hit link to an article published by the Federal Reserve Bank of Cleveland in February 1996. And SavvySearch, AltaVista, Google and Direct Hit all link to an article on Glass-

Steagall and investment banking that was posted in April 1998; the latter two had it at the top of the link list.

Clearly, though they might be of some use for general background information, none of those provided any information on the recently passed repeal. Aged links were a problem across the board on this search. Direct Hit, for example, had a link to a statement by Rep. Jim Leach, chairman of the House Committee on Banking and Financial Services, but it dated to June 1996. Indeed, both Direct Hit and AltaVista each provided a link to a site that no longer existed. About the only link that was reasonably up-to-date -- provided by SavvySearch -- was also of dubious value: something of an antireform rant by an insurance agent upset that banks might be getting into his business.

Search 3

Dreamcast

For the final search, we wanted to find out about Sega Enterprises Ltd.'s new game system, Dreamcast. A search at Direct Hit yielded links to several quality Dreamcast sites, with news, reviews and ways to buy online, as well as prominent links to related searches, such as "Dreamcast Chat." AltaVista also came up with solid results, including a link near the top to Sega's corporate Dreamcast page, but the results weren't quite as deep as Direct Hit's. Results from GoTo.com, which opens up spots in its search results for bidding, were unsurprisingly commerce-related -- how to buy Dreamcast consoles and games -- and lacking in top-notch content sites.

SavvySearch and Google had all the top sites in their results, but About.com's showing was disappointing: It had links to Dreamcast-related articles, but several were very old -- and the search didn't offer a quick, high-ranking route to its best offering, its own "Dreamcast Cheats, Codes, and FAQs" page.

So, after three searches, how do the six search engines rate? GoTo.com and About.com produced the most varied results -- sometimes churning out relevant, focused links, other times falling well short of the technology-based engines. The quality of results at AltaVista, Google and Direct Hit was more consistent; in fact, the sites often produced results that closely matched one another's. SavvySearch was in between, producing a more eclectic batch of links.

Mr. Hanrahan is an editor for The Wall Street Journal Interactive Edition in New York.

INQUERY AND RELEVANCE-RANKING

The THOMAS system uses InQuery search software developed at the University of Massachusetts. InQuery employs a relevance-ranking algorithm to return documents and displays the most relevant documents at the top of the search results list. Documents whose content matches the search terms(s) are retrieved and assigned a "weight" based on an algorithm developed by InQuery programmers.

In general, InQuery calculates the weight of each term for a given document by dividing the number of times the term appears in the document (term frequency) by the number of documents in which the term appears (inverse document frequency) -- that is, the "uniqueness" of the term in the entire database is considered. In a legislative database, the word "Act" is not a unique term, would appear in many documents, and thus be accorded a lower weighting factor than words occurring fewer times in the database, but more times in an individual document. A factor is also added to compensate for the size of the document. Thus, a short document containing 10 instances of a term will be given a higher weight than a much longer document with 10 instances of the same search term. The weights of all the terms in the search statement are then averaged to give an overall weight to the document and rank it in the search results. (The 250 or so stopwords maintained in a THOMAS stopword list are not weighted and are not reflected in search results.)

Words entered in the search box are weighted according to the following criteria (in every case, the uniqueness of the word in the database is also factored in -- inverse document frequency -- as well as the number of times the word(s) occur in relationship to the length of the document):

- SINGLE-WORD SEARCHES

  If only **one** search term is entered, the more instances of that word in the document, the more relevant the document will be considered. Documents with the occurrence of the search term in the title will be considered **most** relevant.

- MULTIPLE-WORD SEARCHES Ranked in Order of Relevance

  1. If more than one word is entered for the search, documents containing instances of those words as a **phrase**--that is, adjacent to each other (discounting "stopwords" or "noisewords") in the order entered-- are considered most relevant. Documents having the occurrence of the search phrase in the **title** are given additional weight.

  2. When more than one word is entered, and the words occur near, but not next to, each other (for example, within a "window" of 30 words), and not necessarily in the same order as entered, the document is ranked less relevant than if the words occur as an exact phrase but more relevant than if the words occur singly, with no proximity.

  3. Documents of yet lesser relevance are those in which all words entered appear singly, not in proximity to each other.

  4. Documents of least relevance will be those which, if more than one word is entered, contain the occurrence of less than all of the words.

  5. Documents which InQuery considers of NO relevance -- containing NO instances of any form of the search words -- will not appear on the INQUERY results list, even though there might be bills which the searcher considers germane. For example, if the searcher enters the query: `capital punishment`, but the document speaks not of **capital punishment**, but instead ONLY of the **death penalty**, that document

will not be returned, even though it is well within the scope of the user's intended search. Future refinements of InQuery search algorithms, employing a legislative thesaurus, may overcome this problem.

The searcher who wishes to use offical subject terms (index terms) to overcome this problem may often identify a more complete set of relevant bills by searching in the THOMAS Bill Summary & Status files, using the searching by subject term option. Records for bills returned in that search will have links back to the full text of the bill in the Bill Text files.

THE RANKED RESULTS OF A SAMPLE SEARCH

Consider the search:

```
defense appropriations
```

Documents with the exact phrase "defense appropriations" appearing in the title or numerous times in the text will appear at the top of the results list -- i.e., they are considered most relevant.

Documents which DO NOT contain the phrase "defense appropriations," but a phrase such as "monies appropriated for the staff of the Department of Defense" -- which contains the words, or variants of the word (i.e., *appropriated*) but not next to each other, as an exact phrase -- will rank lower on the list.

Documents which contain any form of the word "defense" and any form of the word "appropriations" (but not near each other, as explained above) will rank even lower on the results list.

Finally, not all documents in the results set will necessarily contain BOTH words. A document having occurrences of EITHER a form of the word "defense" OR a form of the word "appropriations" will appear lowest on the list. Documents with neither any form of the word "defense" nor any form of the word "appropriations" will not appear on the list, even though the searcher might consider them relevant. For example, a document containing numerous instances of the phrases "monies set aside for the military purposes" or "funds supporting the Army, Navy, and Marines" but neither the word "defense" nor "appropriations," will not appear on the InQuery results list, but may be within the scope of the user's intended search. The searcher may improve search results by using synonyms in another search query, or use a subject (index) term search in the THOMAS Bill Summary & Status files.

Since the default number of results (the "hit" list) is set at 100, ALL of the documents which meet all of the above criteria may not be displayed. (The searcher may adjust the maximum number of bills to be retrieved as high as 2000.) However, those displayed will be the more germane to the search query than those not displayed.

For best results, the searcher should use the most unique search words possible to express his/her concept, avoiding common words which are likely to appear frequently in the database -- e.g., example, "bill" or "act" or "Congress."

For example, the search

```
brady handgun
```

is preferable to the search

```
brady bill
```

or simply

brady

Homepage | Feedback | About THOMAS

Last Update: Tuesday, June 19, 2001

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page
{sergey, page}@cs.stanford.edu
Computer Science Department, Stanford University, Stanford, CA 94305

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date.

Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords: World Wide Web, Search Engines, Information Retrieval, PageRank, Google

1. Introduction

(Note: There are two versions of this paper -- a longer full version and a shorter printed version. The full version is available on the web and the conference CD-ROM.)
The web creates new challenges for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. People are likely to surf the web using its link graph, often starting with high quality human maintained indices such as Yahoo! or with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead automated search engines. We have built a large-scale search engine which addresses many of the problems of existing systems. It makes especially heavy use of the additional structure present in hypertext to provide much higher quality search results. We chose our system name, Google, because it is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search engines.

1.1 Web Search Engines -- Scaling Up: 1994 - 2000

Search engine technology has had to scale dramatically to keep up with the growth of the web. In 1994, one of the first web search engines, the World Wide Web Worm (WWWW) [McBryan 94] had an index of 110,000 web pages and web accessible documents. As of November, 1997, the top search engines claim to index from 2 million (WebCrawler) to 100 million web documents (from Search Engine Watch). It is foreseeable that by the year 2000, a comprehensive index of the Web will contain over a

billion documents. At the same time, the number of queries search engines handle has grown incredibly too. In March and April 1994, the World Wide Web Worm received an average of about 1500 queries per day. In November 1997, Altavista claimed it handled roughly 20 million queries per day. With the increasing number of users on the web, and automated systems which query search engines, it is likely that top search engines will handle hundreds of millions of queries per day by the year 2000. The goal of our system is to address many of the problems, both in quality and scalability, introduced by scaling search engine technology to such extraordinary numbers.

1.2. Google: Scaling with the Web

Creating a search engine which scales even to today's web presents many challenges. Fast crawling technology is needed to gather the web documents and keep them up to date. Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently. Queries must be handled quickly, at a rate of hundreds to thousands per second.

These tasks are becoming increasingly difficult as the Web grows. However, hardware performance and cost have improved dramatically to partially offset the difficulty. There are, however, several notable exceptions to this progress such as disk seek time and operating system robustness. In designing Google, we have considered both the rate of growth of the Web and technological changes. Google is designed to scale well to extremely large data sets. It makes efficient use of storage space to store the index. Its data structures are optimized for fast and efficient access (see section 4.2). Further, we expect that the cost to index and store text or HTML will eventually decline relative to the amount that will be available (see Appendix B). This will result in favorable scaling properties for centralized systems like Google.

1.3 Design Goals

1.3.1 Improved Search Quality

Our main goal is to improve the quality of web search engines. In 1994, some people believed that a complete search index would make it possible to find anything easily. According to Best of the Web 1994 -- Navigators, "The best navigation service should make it easy to find almost anything on the Web (once all the data is entered)." However, the Web of 1997 is quite different. Anyone who has used a search engine recently, can readily testify that the completeness of the index is not the only factor in the quality of search results. "Junk results" often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results). One of the main causes of this problem is that the number of documents in the indices has been increasing by many orders of magnitude, but the user's ability to look at documents has not. People are still only willing to look at the first few tens of results. Because of this, as the collection size grows, we need tools that have very high precision (number of relevant documents returned, say in the top tens of results). Indeed, we want our notion of "relevant" to only include the very best documents since there may be tens of thousands of slightly relevant documents. This very high precision is important even at the expense of recall (the total number of relevant documents the system is able to return). There is quite a bit of recent optimism that the use of more hypertextual information can help improve search and other applications [Marchiori 97] [Spertus 97] [Weiss 96] [Kleinberg 98]. In particular, link structure [Page 98] and link text provide a lot of information for making relevance judgments and quality filtering. Google makes use of both link structure and anchor text (see Sections 2.1 and 2.2).

1.3.2 Academic Search Engine Research

Aside from tremendous growth, the Web has also become increasingly commercial over time. In 1993, 1.5% of web servers were on .com domains. This number grew to over 60% in 1997. At the same time, search engines have migrated from the academic domain to the commercial. Up until now most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art and to be advertising oriented (see Appendix A). With Google, we have a strong goal to push more development and understanding into the academic

'realm.

Another important design goal was to build systems that reasonable numbers of people can actually use. Usage was important to us because we think some of the most interesting research will involve leveraging the vast amount of usage data that is available from modern web systems. For example, there are many tens of millions of searches performed every day. However, it is very difficult to get this data, mainly because it is considered commercially valuable.

Our final design goal was to build an architecture that can support novel research activities on large-scale web data. To support novel research uses, Google stores all of the actual documents it crawls in compressed form. One of our main goals in designing Google was to set up an environment where other researchers can come in quickly, process large chunks of the web, and produce interesting results that would have been very difficult to produce otherwise. In the short time the system has been up, there have already been several papers using databases generated by Google, and many others are underway. Another goal we have is to set up a Spacelab-like environment where researchers or even students can propose and do interesting experiments on our large-scale web data.

2. System Features

The Google search engine has two important features that help it produce high precision results. First, it makes use of the link structure of the Web to calculate a quality ranking for each web page. This ranking is called PageRank and is described in detail in [Page 98]. Second, Google utilizes link to improve search results.

2.1 PageRank: Bringing Order to the Web

The citation (link) graph of the web is an important resource that has largely gone unused in existing web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow rapid calculation of a web page's "PageRank", an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, PageRank is an excellent way to prioritize the results of web keyword searches. For most popular subjects, a simple text matching search that is restricted to web page titles performs admirably when PageRank prioritizes the results (demo available at google.stanford.edu). For the type of full text searches in the main Google system, PageRank also helps a great deal.

2.1.1 Description of PageRank Calculation

Academic citation literature has been applied to the web, largely by counting citations or backlinks to a given page. This gives some approximation of a page's importance or quality. PageRank extends this idea by not counting links from all pages equally, and by normalizing by the number of links on a page. PageRank is defined as follows:

> *We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:*
>
> $$PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$
>
> *Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.*

PageRank or *PR(A)* can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. Also, a PageRank for 26 million web pages can be computed in a few hours on a medium size workstation. There are many other details which are beyond

the scope of this paper.

2.1.2 Intuitive Justification

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the d damping factor is the probability at each page the "random surfer" will get bored and request another random page. One important variation is to only add the damping factor d to a single page, or a group of pages. This allows for personalization and can make it nearly impossible to deliberately mislead the system in order to get a higher ranking. We have several other extensions to PageRank, again see [Page 98].

Another intuitive justification is that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Intuitively, pages that are well cited from many places around the web are worth looking at. Also, pages that have perhaps only one citation from something like the Yahoo! homepage are also generally worth looking at. If a page was not high quality, or was a broken link, it is quite likely that Yahoo's homepage would not link to it. PageRank handles both these cases and everything in between by recursively propagating weights through the link structure of the web.

2.2 Anchor Text

The text of links is treated in a special way in our search engine. Most search engines associate the text of a link with the page that the link is on. In addition, we associate it with the page the link points to. This has several advantages. First, anchors often provide more accurate descriptions of web pages than the pages themselves. Second, anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programs, and databases. This makes it possible to return web pages which have not actually been crawled. Note that pages that have not been crawled can cause problems, since they are never checked for validity before being returned to the user. In this case, the search engine can even return a page that never actually existed, but had hyperlinks pointing to it. However, it is possible to sort the results, so that this particular problem rarely happens.

This idea of propagating anchor text to the page it refers to was implemented in the World Wide Web Worm [McBryan 94] especially because it helps search non-text information, and expands the search coverage with fewer downloaded documents. We use anchor propagation mostly because anchor text can help provide better quality results. Using anchor text efficiently is technically difficult because of the large amounts of data which must be processed. In our current crawl of 24 million pages, we had over 259 million anchors which we indexed.

2.3 Other Features

Aside from PageRank and the use of anchor text, Google has several other features. First, it has location information for all hits and so it makes extensive use of proximity in search. Second, Google keeps track of some visual presentation details such as font size of words. Words in a larger or bolder font are weighted higher than other words. Third, full raw HTML of pages is available in a repository.

3 Related Work

Search research on the web has a short and concise history. The World Wide Web Worm (WWWW) [McBryan 94] was one of the first web search engines. It was subsequently followed by several other academic search engines, many of which are now public companies. Compared to the growth of the Web and the importance of search engines there are precious few documents about recent search engines [Pinkerton 94]. According to Michael Mauldin (chief scientist, Lycos Inc) [Mauldin], "the various services (including Lycos) closely guard the details of these databases". However, there has been a fair amount of work on specific features of search engines. Especially well represented is work which can

get results by post-processing the results of existing commercial search engines, or produce small scale "individualized" search engines. Finally, there has been a lot of research on information retrieval systems, especially on well controlled collections. In the next two sections, we discuss some areas where this research needs to be extended to work better on the web.

3.1 Information Retrieval

Work in information retrieval systems goes back many years and is well developed [Witten 94]. However, most of the research on information retrieval systems is on small well controlled homogeneous collections such as collections of scientific papers or news stories on a related topic. Indeed, the primary benchmark for information retrieval, the Text Retrieval Conference [TREC 96], uses a fairly small, well controlled collection for their benchmarks. The "Very Large Corpus" benchmark is only 20GB compared to the 147GB from our crawl of 24 million web pages. Things that work well on TREC often do not produce good results on the web. For example, the standard vector space model tries to return the document that most closely approximates the query, given that both query and document are vectors defined by their word occurrence. On the web, this strategy often returns very short documents that are the query plus a few words. For example, we have seen a major search engine return a page containing only "Bill Clinton Sucks" and picture from a "Bill Clinton" query. Some argue that on the web, users should specify more accurately what they want and add more words to their query. We disagree vehemently with this position. If a user issues a query like "Bill Clinton" they should get reasonable results since there is a enormous amount of high quality information available on this topic. Given examples like these, we believe that the standard information retrieval work needs to be extended to deal effectively with the web.

3.2 Differences Between the Web and Well Controlled Collections

The web is a vast collection of completely uncontrolled heterogeneous documents. Documents on the web have extreme variation internal to the documents, and also in the external meta information that might be available. For example, documents differ internally in their language (both human and programming), vocabulary (email addresses, links, zip codes, phone numbers, product numbers), type or format (text, HTML, PDF, images, sounds), and may even be machine generated (log files or output from a database). On the other hand, we define external meta information as information that can be inferred about a document, but is not contained within it. Examples of external meta information include things like reputation of the source, update frequency, quality, popularity or usage, and citations. Not only are the possible sources of external meta information varied, but the things that are being measured vary many orders of magnitude as well. For example, compare the usage information from a major homepage, like Yahoo's which currently receives millions of page views every day with an obscure historical article which might receive one view every ten years. Clearly, these two items must be treated very differently by a search engine.

Another big difference between the web and traditional well controlled collections is that there is virtually no control over what people can put on the web. Couple this flexibility to publish anything with the enormous influence of search engines to route traffic and companies which deliberately manipulating search engines for profit become a serious problem. This problem that has not been addressed in traditional closed information retrieval systems. Also, it is interesting to note that metadata efforts have largely failed with web search engines, because any text on the page which is not directly represented to the user is abused to manipulate search engines. There are even numerous companies which specialize in manipulating search engines for profit.

4 System Anatomy

First, we will provide a high level discussion of the architecture. Then, there is some in-depth descriptions of important data structures. Finally, the major applications: crawling, indexing, and searching will be examined in depth.

BEST AVAILABLE COPY

4.1 Google Architecture Overview

In this section, we will give a high level overview of how the whole system works as pictured in Figure 1. Further sections will discuss the applications and data structures not mentioned in this section. Most of Google is implemented in C or C++ for efficiency and can run in either Solaris or Linux.

In Google, the web crawling (downloading of web pages) is done by several distributed crawlers. There is a URLserver that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the storeserver. The storeserver then compresses and stores the web pages into a repository. Every web page has an associated ID number called a docID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions. It reads the repository, uncompresses the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an

Figure 1. High Level Google Architecture

approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link.

The URLresolver reads the anchors file and converts relative URLs into absolute URLs and in turn into docIDs. It puts the anchor text into the forward index, associated with the docID that the anchor points to. It also generates a database of links which are pairs of docIDs. The links database is used to compute PageRanks for all the documents.

The sorter takes the barrels, which are sorted by docID (this is a simplification, see Section 4.2.5), and resorts them by wordID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of wordIDs and offsets into the inverted index. A program called DumpLexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used by the searcher. The searcher is run by a web server and uses the lexicon built by DumpLexicon together with the inverted index and the PageRanks to answer queries.

4.2 Major Data Structures

Google's data structures are optimized so that a large document collection can be crawled, indexed, and searched with little cost. Although, CPUs and bulk input output rates have improved dramatically over the years, a disk seek still requires about 10 ms to complete. Google is designed to avoid disk seeks whenever possible, and this has had a considerable influence on the design of the data structures.

4.2.1 BigFiles

BigFiles are virtual files spanning multiple file systems and are addressable by 64 bit integers. The allocation among multiple file systems is handled automatically. The BigFiles package also handles allocation and deallocation of file descriptors, since the operating systems do not provide enough for our needs. BigFiles also support rudimentary compression options.

4.2.2 Repository

The repository contains the full HTML of every web page. Each page is compressed using zlib (see RFC1950). The choice of compression technique is a tradeoff between speed and compression ratio. We chose zlib's speed over a significant improvement in compression offered by bzip. The compression rate of bzip was approximately 4 to 1 on the repository as compared to zlib's 3 to 1 compression. In the repository, the documents are stored one after the other and are prefixed by docID, length, and URL as can be seen in Figure 2. The repository requires no other data structures to be used in order to access it. This helps with data consistency and makes development much easier; we can rebuild all the other data structures from only the repository and a file which lists crawler errors.

Repository: 53.5 GB = 147.8 GB uncompressed

sync	length	compressed packet
sync	length	compressed packet

...

Packet (stored compressed in repository)

docid	ecode	urllen	pagelen	url	page

Figure 2. Repository Data Structure

4.2.3 Document Index

The document index keeps information about each document. It is a fixed width ISAM (Index sequential access mode) index, ordered by docID. The information stored in each entry includes the current document status, a pointer into the repository, a document checksum, and various statistics. If the document has been crawled, it also contains a pointer into a variable width file called docinfo which contains its URL and title. Otherwise the pointer points into the URLlist which contains just the URL. This design decision was driven by the desire to have a reasonably compact data structure, and the ability to fetch a record in one disk seek during a search

Additionally, there is a file which is used to convert URLs into docIDs. It is a list of URL checksums with their corresponding docIDs and is sorted by checksum. In order to find the docID of a particular URL, the URL's checksum is computed and a binary search is performed on the checksums file to find its docID. URLs may be converted into docIDs in batch by doing a merge with this file. This is the technique the URLresolver uses to turn URLs into docIDs. This batch mode of update is crucial because otherwise we must perform one seek for every link which assuming one disk would take more than a month for our 322 million link dataset.

4.2.4 Lexicon

The lexicon has several different forms. One important change from earlier systems is that the lexicon can fit in memory for a reasonable price. In the current implementation we can keep the lexicon in memory on a machine with 256 MB of main memory. The current lexicon contains 14 million words (though some rare words were not added to the lexicon). It is implemented in two parts -- a list of the words (concatenated together but separated by nulls) and a hash table of pointers. For various functions, the list of words has some auxiliary information which is beyond the scope of this paper to explain fully.

4.2.5 Hit Lists

A hit list corresponds to a list of occurrences of a particular word in a particular document including position, font, and capitalization information. Hit lists account for most of the space used in both the forward and the inverted indices. Because of this, it is important to represent them as efficiently as possible. We considered several alternatives for encoding position, font, and capitalization -- simple encoding (a triple of integers), a compact encoding (a hand optimized allocation of bits), and Huffman coding. In the end we chose a hand optimized compact encoding since it required far less space than the simple encoding and far less bit manipulation than Huffman coding. The details of the hits are shown in Figure 3.

Our compact encoding uses two bytes for every hit. There are two types of hits: fancy hits and plain hits. Fancy hits include hits occurring in a URL, title, anchor text, or meta tag. Plain hits include everything else. A plain hit consists of a capitalization bit, font size, and 12 bits of word position in a document (all

positions higher than 4095 are labeled 4096). Font size is represented relative to the rest of the document using three bits (only 7 values are actually used because 111 is the flag that signals a fancy hit). A fancy hit consists of a capitalization bit, the font size set to 7 to indicate it is a fancy hit, 4 bits to encode the type of fancy hit, and 8 bits of position. For anchor hits, the 8 bits of position are split into 4 bits for position in anchor and 4 bits for a hash of the docID the anchor occurs in. This gives us some limited phrase searching as long as there are not that many anchors for a particular word. We expect to update the way that anchor hits are stored to allow for greater resolution in the position and docIDhash fields. We use font size relative to the rest of the document because when searching, you do not want to rank otherwise identical documents differently just because one of the documents is in a larger font.

The length of a hit list is stored before the hits themselves. To save space, the length of the hit list is combined with the wordID in the forward index and the docID in the inverted index. This limits it to 8 and 5 bits respectively (there are some tricks which allow 8 bits to be borrowed from the wordID). If the length is longer than would fit in that many bits, an escape code is used in those bits, and the next two bytes contain the actual length.

Hit: 2 bytes

plain:	cap:1	imp:3	position: 12		
fancy:	cap:1	imp = 7	type: 4	position: 8	
anchor:	cap:1	imp = 7	type: 4	hash:4	pos: 4

Forward Barrels: total 43 GB

docid	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	null wordid		
docid	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	null wordid		

...

4.2.6 Forward Index

The forward index is actually already partially sorted. It is stored in a number of barrels (we used 64). Each barrel holds a range of wordID's. If a document contains words that fall into a particular barrel, the docID is recorded into the barrel, followed by a list of wordID's with hitlists which correspond to those words. This scheme requires slightly more storage because of duplicated docIDs but the difference is very small for a reasonable number of buckets and saves considerable time and

Lexicon: 293MB   Inverted Barrels: 41 GB

wordid	ndocs		docid: 27	nhits:5	hit hit hit hit
wordid	ndocs		docid: 27	nhits:5	hit hit hit
wordid	ndocs		docid: 27	nhits:5	hit hit hit hit
			docid: 27	nhits:5	hit hit

...

Figure 3. Forward and Reverse Indexes and the Lexicon

coding complexity in the final indexing phase done by the sorter. Furthermore, instead of storing actual wordID's, we store each wordID as a relative difference from the minimum wordID that falls into the barrel the wordID is in. This way, we can use just 24 bits for the wordID's in the unsorted barrels, leaving 8 bits for the hit list length.

4.2.7 Inverted Index

The inverted index consists of the same barrels as the forward index, except that they have been processed by the sorter. For every valid wordID, the lexicon contains a pointer into the barrel that wordID falls into. It points to a doclist of docID's together with their corresponding hit lists. This doclist represents all the occurrences of that word in all documents.

An important issue is in what order the docID's should appear in the doclist. One simple solution is to store them sorted by docID. This allows for quick merging of different doclists for multiple word queries. Another option is to store them sorted by a ranking of the occurrence of the word in each document. This makes answering one word queries trivial and makes it likely that the answers to multiple word queries are near the start. However, merging is much more difficult. Also, this makes development much more difficult in that a change to the ranking function requires a rebuild of the index. We chose a compromise between these options, keeping two sets of inverted barrels -- one set for hit lists which include title or anchor hits and another set for all hit lists. This way, we check the first set of barrels first and if there are not enough matches within those barrels we check the larger ones.

4.3 Crawling the Web

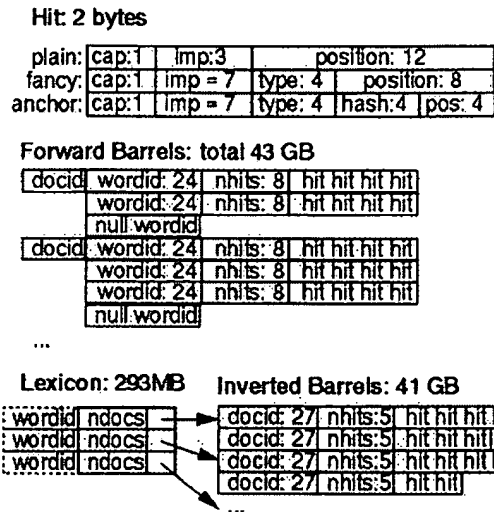Running a web crawler is a challenging task. There are tricky performance and reliability issues and

even more importantly, there are social issues. Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers and various name servers which are all beyond the control of the system.

In order to scale to hundreds of millions of web pages, Google has a fast distributed crawling system. A single URLserver serves lists of URLs to a number of crawlers (we typically ran about 3). Both the URLserver and the crawlers are implemented in Python. Each crawler keeps roughly 300 connections open at once. This is necessary to retrieve web pages at a fast enough pace. At peak speeds, the system can crawl over 100 web pages per second using four crawlers. This amounts to roughly 600K per second of data. A major performance stress is DNS lookup. Each crawler maintains a its own DNS cache so it does not need to do a DNS lookup before crawling each document. Each of the hundreds of connections can be in a number of different states: looking up DNS, connecting to host, sending request, and receiving response. These factors make the crawler a complex component of the system. It uses asynchronous IO to manage events, and a number of queues to move page fetches from state to state.

It turns out that running a crawler which connects to more than half a million servers, and generates tens of millions of log entries generates a fair amount of email and phone calls. Because of the vast number of people coming on line, there are always those who do not know what a crawler is, because this is the first one they have seen. Almost daily, we receive an email something like, "Wow, you looked at a lot of pages from my web site. How did you like it?" There are also some people who do not know about the robots exclusion protocol, and think their page should be protected from indexing by a statement like, "This page is copyrighted and should not be indexed", which needless to say is difficult for web crawlers to understand. Also, because of the huge amount of data involved, unexpected things will happen. For example, our system tried to crawl an online game. This resulted in lots of garbage messages in the middle of their game! It turns out this was an easy problem to fix. But this problem had not come up until we had downloaded tens of millions of pages. Because of the immense variation in web pages and servers, it is virtually impossible to test a crawler without running it on large part of the Internet. Invariably, there are hundreds of obscure problems which may only occur on one page out of the whole web and cause the crawler to crash, or worse, cause unpredictable or incorrect behavior. Systems which access large parts of the Internet need to be designed to be very robust and carefully tested. Since large complex systems such as crawlers will invariably cause problems, there needs to be significant resources devoted to reading the email and solving these problems as they come up.

4.4 Indexing the Web

- **Parsing** — Any parser which is designed to run on the entire Web must handle a huge array of possible errors. These range from typos in HTML tags to kilobytes of zeros in the middle of a tag, non-ASCII characters, HTML tags nested hundreds deep, and a great variety of other errors that challenge anyone's imagination to come up with equally creative ones. For maximum speed, instead of using YACC to generate a CFG parser, we use flex to generate a lexical analyzer which we outfit with its own stack. Developing this parser which runs at a reasonable speed and is very robust involved a fair amount of work.
- **Indexing Documents into Barrels** -- After each document is parsed, it is encoded into a number of barrels. Every word is converted into a wordID by using an in-memory hash table -- the lexicon. New additions to the lexicon hash table are logged to a file. Once the words are converted into wordID's, their occurrences in the current document are translated into hit lists and are written into the forward barrels. The main difficulty with parallelization of the indexing phase is that the lexicon needs to be shared. Instead of sharing the lexicon, we took the approach of writing a log of all the extra words that were not in a base lexicon, which we fixed at 14 million words. That way multiple indexers can run in parallel and then the small log file of extra words can be processed by one final indexer.
- **Sorting** -- In order to generate the inverted index, the sorter takes each of the forward barrels and sorts it by wordID to produce an inverted barrel for title and anchor hits and a full text inverted barrel. This process happens one barrel at a time, thus requiring little temporary storage. Also, we parallelize the sorting phase to use as many machines as we have simply by running multiple sorters, which can process different buckets at the same time. Since the barrels don't fit into main memory, the sorter further subdivides them into baskets which do fit into memory based on

wordID and docID. Then the sorter, loads each basket into memory, sorts it and writes its contents into the short inverted barrel and the full inverted barrel.

4.5 Searching

The goal of searching is to provide quality search results efficiently. Many of the large commercial search engines seemed to have made great progress in terms of efficiency. Therefore, we have focused more on quality of search in our research, although we believe our solutions are scalable to commercial volumes with a bit more effort. The google query evaluation process is show in Figure 4.

To put a limit on response time, once a certain number (currently 40,000) of matching documents are found, the searcher automatically goes to step 8 in Figure 4. This means that it is possible that sub-optimal results would be returned. We are currently investigating other ways to solve this problem. In the past, we sorted the hits according to PageRank, which seemed to improve the situation.

4.5.1 The Ranking System

Google maintains much more information about web documents than typical search engines. Every hitlist includes position, font, and capitalization information. Additionally, we factor in hits from anchor text and the PageRank of the document. Combining all of this information into a rank is difficult. We designed our ranking function so that no particular factor can have too much influence. First, consider the simplest case -- a single word query. In order to rank a document with a single word query, Google looks at that document's hit list for that word. Google considers each hit to be one of several different types (title, anchor, URL, plain text large font, plain text small font, ...), each of which has its own type-weight. The type-weights make up a vector indexed by type. Google counts the number of hits of each type in the hit list. Then every count is converted into a count-weight. Count-weights increase linearly with counts at first but quickly taper off so that more than a certain count will not help. We take the dot product of the vector of count-weights with the vector of type-weights to compute an IR score for the document. Finally, the IR score is combined with PageRank to give a final rank to the document.

```
1. Parse the query.
2. Convert words into wordIDs.
3. Seek to the start of the doclist in
   the short barrel for every word.
4. Scan through the doclists until
   there is a document that matches
   all the search terms.
5. Compute the rank of that
   document for the query.
6. If we are in the short barrels and at
   the end of any doclist, seek to the
   start of the doclist in the full barrel
   for every word and go to step 4.
7. If we are not at the end of any
   doclist go to step 4.
   Sort the documents that have
   matched by rank and return the
   top k.
```

Figure 4. Google Query Evaluation

For a multi-word search, the situation is more complicated. Now multiple hit lists must be scanned through at once so that hits occurring close together in a document are weighted higher than hits occurring far apart. The hits from the multiple hit lists are matched up so that nearby hits are matched together. For every matched set of hits, a proximity is computed. The proximity is based on how far apart the hits are in the document (or anchor) but is classified into 10 different value "bins" ranging from a phrase match to "not even close". Counts are computed not only for every type of hit but for every type and proximity. Every type and proximity pair has a type-prox-weight. The counts are converted into count-weights and we take the dot product of the count-weights and the type-prox-weights to compute an IR score. All of these numbers and matrices can all be displayed with the search results using a special debug mode. These displays have been very helpful in developing the ranking system.

4.5.2 Feedback

The ranking function has many parameters like the type-weights and the type-prox-weights. Figuring out the right values for these parameters is something of a black art. In order to do this, we have a user feedback mechanism in the search engine. A trusted user may optionally evaluate all of the results that are returned. This feedback is saved. Then when we modify the ranking function, we can see the impact of this change on all previous searches which were ranked. Although far from perfect, this gives us some

idea of how a change in the ranking function affects the search results.

5 Results and Performance

The most important measure of a search engine is the quality of its search results. While a complete user evaluation is beyond the scope of this paper, our own experience with Google has shown it to produce better results than the major commercial search engines for most searches. As an example which illustrates the use of PageRank, anchor text, and proximity, Figure 4 shows Google's results for a search on "bill clinton". These results demonstrates some of Google's features. The results are clustered by server. This helps considerably when sifting through result sets. A number of results are from the whitehouse.gov domain which is what one may reasonably expect from such a search. Currently, most major commercial search engines do not return any results from whitehouse.gov, much less the right ones. Notice that there is no title for the first result. This is because it was not crawled. Instead, Google relied on anchor text to determine this was a good answer to the query. Similarly, the fifth result is an email address which, of course, is not crawlable. It is also a result of anchor text.

```
Query: bill clinton
http://www.whitehouse.gov/
100.00% ━━━━ (no date) (OK)
http://www.whitehouse.gov/
        Office of the President
        99.67%━━━━ (Dec 23 1996) (2K)
        http://www.whitehouse.gov/WH/EOP/OP/html/OP_Home.html
        Welcome To The White House
        99.98% ━━━━ (Nov 09 1997) (5K)
        http://www.whitehouse.gov/WH/Welcome.html
        Send Electronic Mail to the President
        99.86% ━━━━ (Jul 14 1997) (5K)
        http://www.whitehouse.gov/WH/Mail/html/Mail_President.html
mailto:president@whitehouse.gov
99.98% ━━━━
        mailto:President@whitehouse.gov
        99.27% ━━━━
The "Unofficial" Bill Clinton
94.06%━━━━ (Nov 11 1997) (14K)
http://zpub.com/un/un-bc.html
        Bill Clinton Meets The Shrinks
        86.27% ━━━━ (Jun 29 1997) (63K)
        http://zpub.com/un/un-bc9.html
President Bill Clinton - The Dark Side
97.27% ━━━━ (Nov 10 1997) (15K)
http://www.realchange.org/clinton.htm
$3 Bill Clinton
94.73% ━━━━ (no date) (4K)
http://www.gatewy.net/~tjohnson/clinton1.html
```

Figure 4. Sample Results from Google

All of the results are reasonably high quality pages and, at last check, none were broken links. This is largely because they all have high PageRank. The PageRanks are the percentages in red along with bar graphs. Finally, there are no results about a Bill other than Clinton or about a Clinton other than Bill. This is because we place heavy importance on the proximity of word occurrences. Of course a true test of the quality of a search engine would involve an extensive user study or results analysis which we do not have room for here. Instead, we invite the reader to try Google for themselves at http://google.stanford.edu.

5.1 Storage Requirements

Aside from search quality, Google is designed to scale cost effectively to the size of the Web as it grows. One aspect of this is to use storage efficiently. Table 1 has a breakdown of some statistics and storage requirements of Google. Due to compression the total size of the repository is about 53 GB, just over one third of the total data it stores. At current disk prices this makes the repository a relatively cheap source of useful data. More importantly, the total of all the data used by the search engine requires a comparable amount of storage, about 55 GB. Furthermore, most queries can be answered using just the short inverted index. With better encoding and compression of the Document Index, a high quality web search engine may fit onto a 7GB drive of a new PC.
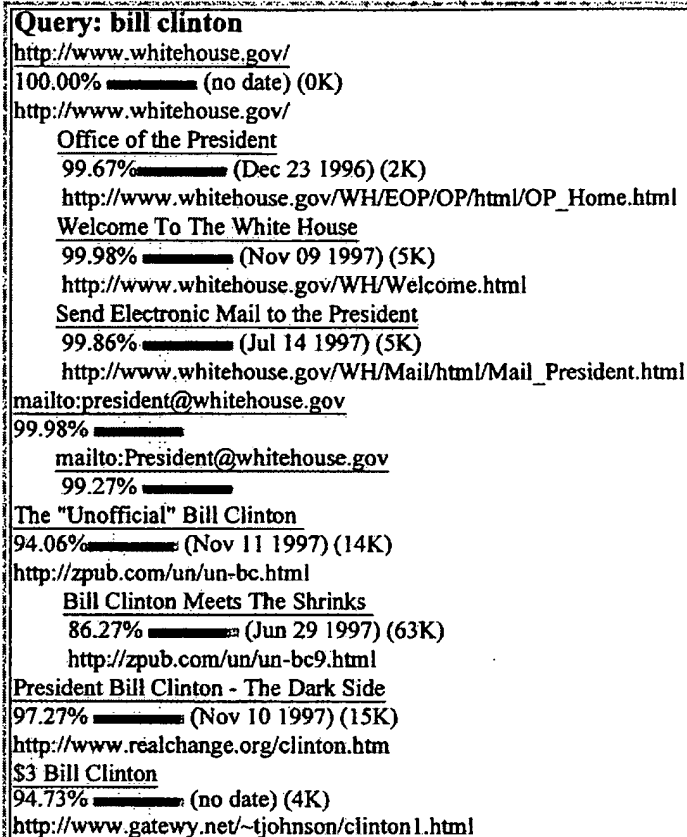
5.2 System Performance

It is important for a search engine to crawl and index efficiently. This way information can be kept up to date and major changes to the system can be tested relatively quickly. For Google, the major operations are Crawling, Indexing, and Sorting. It is difficult to measure how long crawling took overall because disks filled up, name servers crashed, or any number of other problems which stopped the system. In total it took roughly 9 days to download the 26 million pages (including errors). However, once the system was running smoothly, it ran much faster, downloading the last 11 million pages in just 63 hours, averaging just over 4 million pages per day or 48.5 pages per second. We ran the indexer and the crawler simultaneously. The indexer ran just faster than the crawlers. This is largely because we spent just enough time optimizing the indexer so that it would not be a bottleneck. These optimizations included bulk updates to the document index and placement of critical data structures on the local disk. The indexer runs at roughly 54 pages per second. The sorters can be run completely in parallel; using four machines, the whole process of sorting takes about 24 hours.

5.3 Search Performance

Storage Statistics	
Total Size of Fetched Pages	147.8 GB
Compressed Repository	53.5 GB
Short Inverted Index	4.1 GB
Full Inverted Index	37.2 GB
Lexicon	293 MB
Temporary Anchor Data (not in total)	6.6 GB
Document Index Incl. Variable Width Data	9.7 GB
Links Database	3.9 GB
Total Without Repository	55.2 GB
Total With Repository	108.7 GB

Web Page Statistics	
Number of Web Pages Fetched	24 million
Number of Urls Seen	76.5 million
Number of Email Addresses	1.7 million
Number of 404's	1.6 million

Table 1. Statistics

Improving the performance of search was not the major focus of our research up to this point. The current version of Google answers most queries in between 1 and 10 seconds. This time is mostly dominated by disk IO over NFS (since disks are spread over a number of machines). Furthermore, Google does not have any optimizations such as query caching, subindices on common terms, and other common optimizations. We intend to speed up Google considerably through distribution and hardware, software, and algorithmic improvements. Our target is to be able to handle several hundred queries per second. Table 2 has some sample query times from the current version of Google. They are repeated to show the speedups resulting from cached IO.

6 Conclusions

Google is designed to be a scalable search engine. The primary goal is to provide high quality search results over a rapidly growing World Wide Web. Google employs a number of techniques to improve search quality including page rank, anchor text, and proximity information. Furthermore, Google is a complete architecture for gathering web pages, indexing them, and performing search queries over them.

6.1 Future Work

A large-scale web search engine is a complex system and much remains to be done. Our immediate goals are to improve search efficiency and to scale to approximately 100 million web pages. Some simple improvements to efficiency include query

	Initial Query		Same Query Repeated (IO mostly cached)	
Query	CPU Time(s)	Total Time(s)	CPU Time(s)	Total Time(s)
al gore	0.09	2.13	0.06	0.06
vice president	1.77	3.84	1.66	1.80
hard disks	0.25	4.86	0.20	0.24
search engines	1.31	9.63	1.16	1.16

Table 2. Search Times

caching, smart disk allocation, and subindices. Another area which requires much research is updates. We must have smart algorithms to decide what old web pages should be recrawled and what new ones should be crawled. Work toward this goal has been done in [Cho 98]. One promising area of research is using proxy caches to build search databases, since they are demand driven. We are planning to add simple features supported by commercial search engines like boolean operators, negation, and stemming. However, other features are just starting to be explored such as relevance feedback and clustering (Google currently supports a simple hostname based clustering). We also plan to support user context (like the user's location), and result summarization. We are also working to extend the use of link structure and link text. Simple experiments indicate PageRank can be personalized by increasing the weight of a user's home page or bookmarks. As for link text, we are experimenting with using text surrounding links in addition to the link text itself. A Web search engine is a very rich environment for research ideas. We have far too many to list here so we do not expect this Future Work section to become much shorter in the near future.

6.2 High Quality Search

The biggest problem facing users of web search engines today is the quality of the results they get back. While the results are often amusing and expand users' horizons, they are often frustrating and consume precious time. For example, the top result for a search for "Bill Clinton" on one of the most popular commercial search engines was the Bill Clinton Joke of the Day: April 14, 1997. Google is designed to provide higher quality search so as the Web continues to grow rapidly, information can be found easily. In order to accomplish this Google makes heavy use of hypertextual information consisting of link structure and link (anchor) text. Google also uses proximity and font information. While evaluation of a search engine is difficult, we have subjectively found that Google returns higher quality search results than current commercial search engines. The analysis of link structure via PageRank allows Google to evaluate the quality of web pages. The use of link text as a description of what the link points to helps the search engine return relevant (and to some degree high quality) results. Finally, the use of proximity information helps increase relevance a great deal for many queries.

6.3 Scalable Architecture

Aside from the quality of search, Google is designed to scale. It must be efficient in both space and time, and constant factors are very important when dealing with the entire Web. In implementing Google, we have seen bottlenecks in CPU, memory access, memory capacity, disk seeks, disk throughput, disk capacity, and network IO. Google has evolved to overcome a number of these bottlenecks during various operations. Google's major data structures make efficient use of available storage space. Furthermore, the crawling, indexing, and sorting operations are efficient enough to be able to build an index of a substantial portion of the web -- 24 million pages, in less than one week. We expect to be able to build an index of 100 million pages in less than a month.

6.4 A Research Tool

In addition to being a high quality search engine, Google is a research tool. The data Google has collected has already resulted in many other papers submitted to conferences and many more on the way. Recent research such as [Abiteboul 97] has shown a number of limitations to queries about the Web that may be answered without having the Web available locally. This means that Google (or a similar system) is not only a valuable research tool but a necessary one for a wide range of applications. We hope Google will be a resource for searchers and researchers all around the world and will spark the next generation of search engine technology.

7 Acknowledgments

Scott Hassan and Alan Steremberg have been critical to the development of Google. Their talented contributions are irreplaceable, and the authors owe them much gratitude. We would also like to thank Hector Garcia-Molina, Rajeev Motwani, Jeff Ullman, and Terry Winograd and the whole WebBase group for their support and insightful discussions. Finally we would like to recognize the generous

support of our equipment donors IBM, Intel, and Sun and our funders. The research described here was conducted as part of the Stanford Integrated Digital Library Project, supported by the National Science Foundation under Cooperative Agreement IRI-9411306. Funding for this cooperative agreement is also provided by DARPA and NASA, and by Interval Research, and the industrial partners of the Stanford Digital Libraries Project.

References

- Best of the Web 1994 -- Navigators http://botw.org/1994/awards/navigators.html
- Bill Clinton Joke of the Day: April 14, 1997. http://www.io.com/~cjburke/clinton/970414.html.
- Bzip2 Homepage http://www.muraroa.demon.co.uk/
- Google Search Engine http://google.stanford.edu/
- Harvest http://harvest.transarc.com/
- Mauldin, Michael L. Lycos Design Choices in an Internet Search Service, IEEE Expert Interview http://www.computer.org/pubs/expert/1997/trends/x1008/mauldin.htm
- The Effect of Cellular Phone Use Upon Driver Attention http://www.webfirst.com/aaa/text/cell/cell0toc.htm
- Search Engine Watch http://www.searchenginewatch.com/
- RFC 1950 (zlib) ftp://ftp.uu.net/graphics/png/documents/zlib/zdoc-index.html
- Robots Exclusion Protocol: http://info.webcrawler.com/mak/projects/robots/exclusion.htm
- Web Growth Summary: http://www.mit.edu/people/mkgray/net/web-growth-summary.html
- Yahoo! http://www.yahoo.com/

- [Abiteboul 97] Serge Abiteboul and Victor Vianu, *Queries and Computation on the Web*. Proceedings of the International Conference on Database Theory. Delphi, Greece 1997.
- [Bagdikian 97] Ben H. Bagdikian. *The Media Monopoly*. 5th Edition. Publisher: Beacon, ISBN: 0807061557
- [Chakrabarti 98] S.Chakrabarti, B.Dom, D.Gibson, J.Kleinberg, P. Raghavan and S. Rajagopalan. *Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text*. Seventh International Web Conference (WWW 98). Brisbane, Australia, April 14-18, 1998.
- [Cho 98] Junghoo Cho, Hector Garcia-Molina, Lawrence Page. *Efficient Crawling Through URL Ordering*. Seventh International Web Conference (WWW 98). Brisbane, Australia, April 14-18, 1998.
- [Gravano 94] Luis Gravano, Hector Garcia-Molina, and A. Tomasic. *The Effectiveness of GlOSS for the Text-Database Discovery Problem*. Proc. of the 1994 ACM SIGMOD International Conference On Management Of Data, 1994.
- [Kleinberg 98] Jon Kleinberg, *Authoritative Sources in a Hyperlinked Environment*, Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [Marchiori 97] Massimo Marchiori. *The Quest for Correct Information on the Web: Hyper Search Engines*. The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.
- [McBryan 94] Oliver A. McBryan. GENVL and *WWWW: Tools for Taming the Web*. First International Conference on the World Wide Web. CERN, Geneva (Switzerland), May 25-26-27 1994. http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps
- [Page 98] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Manuscript in progress. http://google.stanford.edu/~backrub/pageranksub.ps
- [Pinkerton 94] Brian Pinkerton, *Finding What People Want: Experiences with the WebCrawler*. The Second International WWW Conference Chicago, USA, October 17-20, 1994. http://info.webcrawler.com/bp/WWW94.html
- [Spertus 97] Ellen Spertus. *ParaSite: Mining Structural Information on the Web*. The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.
- [TREC 96] *Proceedings of the fifth Text REtrieval Conference (TREC-5)*. Gaithersburg, Maryland, November 20-22, 1996. Publisher: Department of Commerce, National Institute of Standards and Technology. Editors: D. K. Harman and E. M. Voorhees. Full text at: http://trec.nist.gov/
- [Witten 94] Ian H Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes:*

Compressing and Indexing Documents and Images. New York: Van Nostrand Reinhold, 1994.
- [Weiss 96] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. *HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering*. Proceedings of the 7th ACM Conference on Hypertext. New York, 1996.

Vitae

Sergey Brin received his B.S. degree in mathematics and computer science from the University of Maryland at College Park in 1993. Currently, he is a Ph.D. candidate in computer science at Stanford University where he received his M.S. in 1995. He is a recipient of a National Science Foundation Graduate Fellowship. His research interests include search engines, information extraction from unstructured sources, and data mining of large text collections and scientific data.

Lawrence Page was born in East Lansing, Michigan, and received a B.S.E. in Computer Engineering at the University of Michigan Ann Arbor in 1995. He is currently a Ph.D. candidate in Computer Science at Stanford University. Some of his research interests include the link structure of the web, human computer interaction, search engines, scalability of information access interfaces, and personal data mining.

8 Appendix A: Advertising and Mixed Motives

Currently, the predominant business model for commercial search engines is advertising. The goals of the advertising business model do not always correspond to providing quality search to users. For example, in our prototype search engine one of the top results for cellular phone is "The Effect of Cellular Phone Use Upon Driver Attention", a study which explains in great detail the distractions and risk associated with conversing on a cell phone while driving. This search result came up first because of its high importance as judged by the PageRank algorithm, an approximation of citation importance on the web [Page, 98]. It is clear that a search engine which was taking money for showing cellular phone ads would have difficulty justifying the page that our system returned to its paying advertisers. For this type of reason and historical experience with other media [Bagdikian 83], we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers.

Since it is very difficult even for experts to evaluate search engines, search engine bias is particularly insidious. A good example was OpenText, which was reported to be selling companies the right to be listed at the top of the search results for particular queries [Marchiori 97]. This type of bias is much more insidious than advertising, because it is not clear who "deserves" to be there, and who is willing to pay money to be listed. This business model resulted in an uproar, and OpenText has ceased to be a viable search engine. But less blatant bias are likely to be tolerated by the market. For example, a search engine could add a small factor to search results from "friendly" companies, and subtract a factor from results from competitors. This type of bias is very difficult to detect but could still have a significant effect on the market. Furthermore, advertising income often provides an incentive to provide poor quality search results. For example, we noticed a major search engine would not return a large airline's homepage when the airline's name was given as a query. It so happened that the airline had placed an expensive ad, linked to the query that was its name. A better search engine would not have required this ad, and possibly resulted in the loss of the revenue from the airline to the search engine. In general, it could be argued from the consumer point of view that the better the search engine is, the fewer advertisements will be needed for the consumer to find what they want. This of course erodes the advertising supported

business model of the existing search engines. However, there will always be money from advertisers who want a customer to switch products, or have something that is genuinely new. But we believe the issue of advertising causes enough mixed incentives that it is crucial to have a competitive search engine that is transparent and in the academic realm.

9 Appendix B: Scalability

9.1 Scalability of Google

We have designed Google to be scalable in the near term to a goal of 100 million web pages. We have just received disk and machines to handle roughly that amount. All of the time consuming parts of the system are parallelize and roughly linear time. These include things like the crawlers, indexers, and sorters. We also think that most of the data structures will deal gracefully with the expansion. However, at 100 million web pages we will be very close up against all sorts of operating system limits in the common operating systems (currently we run on both Solaris and Linux). These include things like addressable memory, number of open file descriptors, network sockets and bandwidth, and many others. We believe expanding to a lot more than 100 million pages would greatly increase the complexity of our system.

9.2 Scalability of Centralized Indexing Architectures

As the capabilities of computers increase, it becomes possible to index a very large amount of text for a reasonable cost. Of course, other more bandwidth intensive media such as video is likely to become more pervasive. But, because the cost of production of text is low compared to media like video, text is likely to remain very pervasive. Also, it is likely that soon we will have speech recognition that does a reasonable job converting speech into text, expanding the amount of text available. All of this provides amazing possibilities for centralized indexing. Here is an illustrative example. We assume we want to index everything everyone in the US has written for a year. We assume that there are 250 million people in the US and they write an average of 10k per day. That works out to be about 850 terabytes. Also assume that indexing a terabyte can be done now for a reasonable cost. We also assume that the indexing methods used over the text are linear, or nearly linear in their complexity. Given all these assumptions we can compute how long it would take before we could index our 850 terabytes for a reasonable cost assuming certain growth factors. Moore's Law was defined in 1965 as a doubling every 18 months in processor power. It has held remarkably true, not just for processors, but for other important system parameters such as disk as well. If we assume that Moore's law holds for the future, we need only 10 more doublings, or 15 years to reach our goal of indexing everything everyone in the US has written for a year for a price that a small company could afford. Of course, hardware experts are somewhat concerned Moore's Law may not continue to hold for the next 15 years, but there are certainly a lot of interesting centralized applications even if we only get part of the way to our hypothetical example.

Of course a distributed systems like Gloss [Gravano 94] or Harvest will often be the most efficient and elegant technical solution for indexing, but it seems difficult to convince the world to use these systems because of the high administration costs of setting up large numbers of installations. Of course, it is quite likely that reducing the administration cost drastically is possible. If that happens, and everyone starts running a distributed indexing system, searching would certainly improve drastically.

Because humans can only type or speak a finite amount, and as computers continue improving, text indexing will scale even better than it does now. Of course there could be an infinite amount of machine generated content, but just indexing huge amounts of human generated content seems tremendously useful. So we are optimistic that our centralized web search engine architecture will improve in its ability to cover the pertinent text information over time and that there is a bright future for search.

12/5/1   (Item 1 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

017148828   **Image available**
WPI Acc No: 2005-473173/200548
XRPX Acc No: N05-384692
  **Direct mail management method involves displaying the charge information
  to client and delivering direct mail to mail receiver based on received
  request information selected layer of direct mail received from client
  terminal**
Patent Assignee: NEC CORP (NIDE)
Inventor: SATO Y
Number of Countries: 001 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
JP 2005182261 A   20050707 JP 2003419277   A   20031217 200548 B

Priority Applications (No Type Date): JP 2003419277 A 20031217
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
JP 2005182261 A   15 G06F-017/60

Abstract (Basic): JP 2005182261 A
      NOVELTY - The method involves displaying the charge information to
  a client based on request information specifying selected layer of a
  direct mail received from a client terminal and delivering the direct
  mail to a mail receiver based on the received request information. The
  response information with respect to the direct mail is acquired. The
  receiver **attribute** information is re-categorized based on acquired
  content.
      DETAILED DESCRIPTION - An INDEPENDENT CLAIM is also included for
  direct mail management apparatus.
      USE - For managing direct mail in business **strategy** such as
  market trend investigation and sales promotion through network such as
  internet.
      ADVANTAGE - The direct mail is delivered to the mail receiver
  accurately. The improvement of the reaction **rates** such as clicking
  rate in internet advertising and **site** residence time can be
  anticipated.
      DESCRIPTION OF DRAWING(S) - The figure shows a structure of the
  direct mail management apparatus.(Drawing includes non-English language
  text).
      database (DB) management unit (12)
      DB extraction unit (32)
      data creation unit (33)
      DB change unit (34)
      storage unit (35)
      pp; 15 DwgNo 3/7   -
Title Terms: DIRECT; MAIL; MANAGEMENT; METHOD; DISPLAY; CHARGE; INFORMATION
  ; CLIENT; DELIVER; DIRECT; MAIL; MAIL; RECEIVE; BASED; RECEIVE; REQUEST;
  INFORMATION; SELECT; LAYER; DIRECT; MAIL; RECEIVE; CLIENT; TERMINAL
Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
File Segment: EPI

12/5/2   (Item 2 from file: 350)

14/3,K/1   (Item 1 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

02037792 55143664
Content matters most in search-engine placement
Kahaner, Larry
Informationweek n790 PP: 172-178 Jun 12, 2000
ISSN: 8750-6874 JRNL CODE: IWK
WORD COUNT: 2155

...TEXT: or arcane tactics. It takes Web-design skills, perseverance, hard
work, a thorough knowledge of **how** the various search engines **rank   Web
sites** , a smattering of good Luck, and, most of all, compelling content.

In fact, now that...

14/3,K/2   (Item 2 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

01826036 04-77027
Internet Librarian, LibTech International
Raitt, David
Information Today v16n5 PP: 24-27 May 1999
ISSN: 8755-6286 JRNL CODE: IFT
WORD COUNT: 2843

...TEXT: O'Sullivan, Search Engine Watch, U.K., had interesting things to
say as.well about **how   Web   pages** were **ranked** and **how** to index your
Web pages so that they are found. He suggested making different pages
for different engines since each...

14/3,K/3   (Item 3 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

01812097 04-63088
A Web search trifecta
Mickey, Bill
Online v23n3 PP: 79-82 May/Jun 1999
ISSN: 0146-5422 JRNL CODE: ONL
WORD COUNT: 1442

...TEXT: on the many new search engine developments?

The Webmaster's Guide includes comprehensive discussion on **how** search
engines **rank   Web   pages** and **how** to improve a site's **ranking** . This
section provides an extremely well-rounded introduction to search engines
and how they work...

14/3,K/4   (Item 4 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

01742873   03-93863
Quality of life and economic development policy
Blair, John P
Economic Development Review   v16n1   PP: 50-54   1998
ISSN: 0742-3713   JRNL CODE: EDR
WORD COUNT: 3043

...TEXT: of Money Magazine and Places Rated Almanac, employ
opiniondetermined weights.

Money Magazine has an interesting **web   site** that illustrates **how**
sensitive urban **rankings** can be to quality-of-life weighting. The site
allows browsers to select weights and...

 14/3,K/5   (Item 5 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)
(c) 2005 ProQuest Info&Learning. All rts. reserv.

01693666   03-44656
Net relations: A fusion of direct marketing and public relations
Spataro, Mike
Direct Marketing   v61n4   PP: 16-19   Aug 1998
ISSN: 0012-3188   JRNL CODE: DIM
WORD COUNT: 2108

...TEXT: This is the new paradigm for direct marketing. Extensive analysis
must be conducted to determine **how** a company's **Web   site** is **ranked**
and categorized in major search engines and Web directories. One of the
many challenges is...

 14/3,K/6   (Item 6 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)
(c) 2005 ProQuest Info&Learning. All rts. reserv.

01559644   02-10633
New Year's resolutions for the Web user
Kennedy, Shirley Duglin
Information Today   v15n1   PP: 32, 34   Jan 1998
ISSN: 8755-6286   JRNL CODE: IFT
WORD COUNT: 2223

...TEXT: Web search services. Today, that guide
(http://www.searchenginewatch.com/wgt se.htm)-which "explains **how** search
engines find and **rank   Web   pages** , with an emphasis on what Webmasters
can do to improve how search engines list their...

 14/3,K/7   (Item 1 from file: 610)
DIALOG(R)File 610:Business Wire
(c) 2005 Business Wire. All rts. reserv.

00346350   20000817230B1555   (USE FORMAT 7 FOR FULLTEXT)
Newstream.com digest: Ranking **the Airline** Websites , How **Not to Buy a
Lemon Other Multimedia for Journalists**
Business Wire
Thursday, August 17, 2000 18:29 EDT

JOURNAL CODE: BW LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 221

Newstream.com digest: Ranking **the Airline** Websites , How **Not to Buy a
Lemon Other Multimedia for Journalists**

 14/3,K/8   (Item 1 from file: 810)
DIALOG(R)File 810:Business Wire
(c) 1999 Business Wire . All rts. reserv.

0880106   BW0266

**FIRSTPLACE SOFTWARE: WebPostion Gold 1.0: First Web Site Promotion Software
     To Shift Balance of Power From Search Engines to Marketers**

July 15, 1998

Byline: Business Editors, High-Tech Writers

     ...found near the top of the results, it might
as well be invisible," said Winters. " **How** search engines determine
where a **Web   site   ranks** has long been their most carefully guarded
secret. If a Web marketer is willing to...

 14/3,K/9   (Item 2 from file: 810)
DIALOG(R)File 810:Business Wire
(c) 1999 Business Wire . All rts. reserv.

0827163   BW1251

**SCORECARD: A New Internet Agent, ScoreCheck, Determines Competitive Website
     Rankings On Search Engines In A Volatile Online Advertising Market**

March 26, 1998

Byline: Business Editors/Computers & Electronics Writers

     ...BUSINESS WIRE)--March 26, 1998--ScoreCard Inc.
Thursday announced ScoreCheck, its automated agent that analyzes **how
web   sites   rank** with 15 major internet search engines.
     This service will be an important tool for companies...

 14/3,K/10   (Item 3 from file: 810)
DIALOG(R)File 810:Business Wire
(c) 1999 Business Wire . All rts. reserv.

0729686   BW1208

**IBM SOFTWARE: IBM Software Agent Technology Helps Users Control Web
     Information**

July 30, 1997

Byline: Business & Technology Editors

...the site, provide
alerts to speeds of links, advise the user to changes at a **web   site** ,
 rank order viewed sites by frequency and **how**   recently they have been
visited, learn user patterns and suggest shortcuts.
    For enterprise customers, software...

 14/3,K/11   (Item 1 from file: 476)
DIALOG(R)File 476:Financial Times Fulltext
(c) 2005 Financial Times Ltd. All rts. reserv.

0009019174 B0HFWAGAC6FT
Marketing / Advertising / Media: Never confuse your robot: Tim Jackson . On
  the Web
TIM JACKSON
Financial Times, London Edition 1 ED, P 15
Monday, June 23, 1997
DOCUMENT TYPE: Columns; NEWSPAPER LANGUAGE: ENGLISH   RECORD TYPE:
FULLTEXT
Word Count: 822

...which allows you to type in a search term and the address of your own
website , and find out **how**   highly you **ranked** .

Unfortunately, not everyone can be in the top ten. All advice, no matter
how good...

 14/3,K/12   (Item 1 from file: 613)
DIALOG(R)File 613:PR Newswire
(c) 2005 PR Newswire Association Inc. All rts. reserv.

00368682 20000710SFM049 (USE FORMAT 7 FOR FULLTEXT)
Etown.Com Ranked No. 1 Overall Online Electronics Store
PR Newswire
Monday, July 10, 2000 09:02 EDT
JOURNAL CODE: PR LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 748

TEXT:
...a row that etown.com has ranked in the top two places.

    Gomez reviewed and **ranked** 19 **Web   sites** by **how**   well they served
each of
several customer profiles and by categories such as Customer Confidence...

 14/3,K/13   (Item 2 from file: 613)
DIALOG(R)File 613:PR Newswire
(c) 2005 PR Newswire Association Inc. All rts. reserv.

00324481 20000502LATU026 (USE FORMAT 7 FOR FULLTEXT)
Search Engine Placement Service Launched; Webseed.Com Uses Relevant Text to
Boost Search Engine Rankings
PR Newswire
Tuesday, May 2, 2000 07:01 EDT

JOURNAL CODE: PR LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 297

...a progress graph and activity log showing the steps that have been
completed for their **website** .   **Ranking** analysis reports show exactly
how many
top-ten positions have been achieved during the WebSeed process.

    For more information, visit...

  14/3,K/14   (Item 3 from file: 613)
DIALOG(R)File 613:PR Newswire
(c) 2005 PR Newswire Association Inc. All rts. reserv.

00255755 20000131CLM009   (USE FORMAT 7 FOR FULLTEXT)
Upright Communications Unveils Custom Search Engine Marketing Program
PR Newswire
Monday, January 31, 2000 09:08 EST
JOURNAL CODE: PR LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 451

...list of the most
appropriate key words and phrases, which plays an integral role in **how**
search
engines index and **rank   web   sites** .

    In many cases, users will not take time to view search results beyond
the
third...

  14/3,K/15   (Item 1 from file: 813)
DIALOG(R)File 813:PR Newswire
(c) 1999 PR Newswire Association Inc. All rts. reserv.

1163275   NEM042
**Submit It! Acquires PositionAgent from NetGambit, Delivering Significant
    Competitive Advantage to Customers**

DATE: October 6, 1997   13:25 EDT   WORD COUNT: 609

...our ongoing effort to aggressively meet that need," he said.

    "Now customers can easily check **how** their **Web   sites   rank** and
implement strategies for improving their rankings -- all from the same
vendor," said Younker.   "This...

  14/3,K/16   (Item 1 from file: 634)
DIALOG(R)File 634:San Jose Mercury
(c) 2005 San Jose Mercury News. All rts. reserv.

10526017
ASK JEEVES TO BUY DIRECT HIT
San Jose Mercury News (SJ) - Wednesday, January 26, 2000

By: Compiled from staff and wire reports.
Edition: Morning Final   Section: Business   Page: 1C
Word Count: 71

TEXT:
... 12 percent of its shares outstanding, for Direct Hit. Direct Hit, based
in Wellesley, Mass., **ranks** **Web** **sites** according to **how** much time
searchers spend at them.   ...

14/3,K/17   (Item 2 from file: 634)
DIALOG(R)File 634:San Jose Mercury
(c) 2005 San Jose Mercury News. All rts. reserv.

10228016
**THE NEW WORLD OF SEARCH ENGINES PRESSURE TO MAKE MONEY MEANS CLUTTER,
    CONFUSION FOR CONSUMERS**
San Jose Mercury News (SJ) - Monday, August 16, 1999
By: MONUA JANAH, MERCURY NEWS STAFF WRITER
Edition: Morning Final   Section: Business Monday   Page: 1E
Word Count: 1,356

TEXT:
...Home.
    What was once a simple concept -- using search and index technology to
sort through **Web**   **sites** , then **ranking** them by **how** useful they're
likely to be to the computer user -- has gained a commercial edge...

10/5/1   (Item 1 from file: 2)
DIALOG(R)File   2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

6944005   INSPEC Abstract Number: C2001-07-7250R-009
  Title: Extracting information from the Web for concept learning and collaborative filtering
  Author(s): Cohen, W.W.
  Author Affiliation: WhizBang! Labs.-Res., Pittsburgh, PA, USA
  Conference   Title: Algorithmic   Learning   Theory.   11th   International Conference, ALT 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1968)   p.1-12
  Editor(s): Arimura, H.; Jain, S.; Sharma, A.
  Publisher: Springer-Verlag, Berlin, Germany
  Publication Date: 2000 Country of Publication: Germany   xi+333 pp.
  ISBN: 3 540 41237 9   Material Identity Number: XX-2001-00247
  Conference   Title: Algorithmic   Learning   Theory.   11th   International Conference, ALT 2000. Proceedings
  Conference Sponsor: Univ. New South Wales
  Conference   Date: 11-13   Dec.   2000   Conference Location: Sydney, NSW, Australia
  Language: English   Document Type: Conference Paper (PA)
  Treatment: Theoretical (T); Experimental (X)
  Abstract: Previous work on extracting information from the Web generally makes few assumptions about how the extracted information will be used. As a consequence, the goal of Web-based extraction systems is usually taken to be the creation of high-quality, noise-free data with clear semantics. This is a difficult problem which cannot be completely automated. Here we consider instead the problem of extracting Web data for certain machine learning systems: specifically, collaborative filtering (CF) and concept learning (CL) systems. CF and CL systems are highly tolerant of noisy input, and hence much simpler extraction systems can be used in this context. For CL, we describe a simple **method** that uses a given set of **Web pages** to construct new **features** , which reduce the error **rate** of learned classifiers in a wide variety of situations. For CF, we describe a simple **method** that automatically collects useful information from the Web without any human intervention. The collected information, represented as "pseudo-users", can be used to "jumpstart" a CF system when the user base is small (or even absent). (22 Refs)
  Subfile: C
  Descriptors: data mining; information retrieval; Internet; learning (**artificial intelligence**); pattern classification
  Identifiers: information retrieval; Web data; concept learning; collaborative filtering; pattern classifiers; learning systems
  Class Codes: C7250R (Information retrieval techniques); C1230L (Learning in AI); C6170K (Knowledge engineering techniques); C7210N (Information networks); C1250 (Pattern recognition)
  Copyright 2001, IEE

10/5/2   (Item 2 from file: 2)
DIALOG(R)File   2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

5949314   INSPEC Abstract Number: C9808-7210-020
  Title: What is a tall poppy among Web pages? [WWW information retrieval]
  Author(s): Pringle, G.; Allison, L.; Dowe, D.L.
  Author Affiliation: Sch. of Comput. Sci. & Software Eng., Monash Univ., Clayton, Vic., Australia

Journal: Computer Networks and ISDN Systems Conference Title: Comput. Netw. ISDN Syst. (Netherlands)   vol.30, no.1-7   p.369-77
Publisher: Elsevier,
Publication Date: April 1998 Country of Publication: Netherlands
CODEN: CNISE9 ISSN: 0169-7552
SICI: 0169-7552(199804)30:1/7L.369:WTPA;1-F
Material Identity Number: I876-98002
U.S. Copyright Clearance Center Code: 0169-7552/98/$19.00
Conference Title: 7th International World Wide Web Conference
Conference Date: 14-18 April 1998   Conference Location: Brisbane, Qld., Australia
Document Number: S0169-7552(98)00061-0
Language: English   Document Type: Conference Paper (PA); Journal Paper (JP)
Treatment: Practical (P)

Abstract: Search engines and indices were created to help people find information amongst the rapidly increasing number of World-Wide Web (WWW) pages. The search engines automatically visit and index pages so that they can return good matches for their users' queries. The way that this indexing is done varies from engine to engine and the detail is usually secret although the **strategy** is sometimes made public in general terms. The search engines' aim is to return relevant pages quickly. On the other hand, the author of a **Web page** has a vested interest in it **rating** highly, for appropriate queries, on as many search engines as possible. Some authors have an interest in their **page rating** well for a great many types of query indeed: spamming has come to the Web. We treat the modelling of the workings of WWW search engines as an inductive inference problem. A training set of data is collected, being pages returned in response to typical queries. Decision trees are used as the model class for the search engines' selection **criteria** although this is not to say that search engines actually contain decision trees. A machine learning program is used to infer a decision tree for each search engine, an information-theory **criterion** being used to direct the inference and to prevent over-fitting. (4 Refs)
Subfile: C
Descriptors: inference mechanisms; Internet; learning (**artificial intelligence**); online front-ends; query processing; relevance feedback
Identifiers: search engines; indices; information retrieval; World-Wide Web; WWW pages; relevant pages; inductive inference problem; training set; data collection; queries; decision trees; selection **criteria** ; machine learning program; information-theory **criterion**
Class Codes: C7210 (Information services and centres); C7250N (Front end systems for online searching); C7250R (Information retrieval techniques); C6170K (Knowledge engineering techniques)
Copyright 1998, IEE

10/5/12   (Item 1 from file: 6)
DIALOG(R)File   6:NTIS
(c) 2005 NTIS, Intl Cpyrght All Rights Res. All rts. reserv.

1701733 NTIS Accession Number: DE92041177
   Building performance analysis using interactive multimedia concepts
   Selkowitz, S. ; Beltran, L. ; Osterhaus, W. ; Papamichael, K. ; Schuman, J.
   Lawrence Berkeley Lab., CA.
   Corp. Source Codes: 086929000; 9513034
   Sponsor: Department of Energy, Washington, DC.
   Report No.: LBL-32257; CONF-920828-16

Apr 92   8p
Languages: English   Document Type: Conference proceeding
Journal Announcement: GRAI9307; ERA9310
American Council for an Energy-Efficient Economy (ACEEE) summer study on
energy efficiency in buildings, Pacific Grove, CA (United States), 30 Aug –
5 Sep 1992. Sponsored by Department of Energy, Washington, DC.
Order   this   product   from NTIS by: phone at 1-800-553-NTIS   (U.S.
customers); (703)605-6000 (other countries); fax at (703)321-8547; and
email at orders@ntis.fedworld.gov. NTIS is located at 5285 Port Royal Road,
Springfield, VA, 22161, USA.
NTIS Prices: PC A02/MF A01
Country of Publication: United States
Contract No.: AC03-76SF00098
We describe LBL's involvement with multimedia concepts by discussing
several modules of an advanced computer-based building envelope design
tool. The qualitative and quantitative **aspects** of the building design
process   are accommodated within the same design tool which uses
object-oriented programming **procedures** . This computer-based concept
utilizes images (buildings, landscapes, models, documents, etc.), **expert**
systems   (knowledge bases, i.e., lighting design, **site** planning, HVAC
design, etc.), and data bases (design **criteria** , utility **rates** , climatic
data, etc.) in addition to more traditional simulation models to evaluate
building design alternatives.
Descriptors: *Buildings; *Energy Efficiency; Computer-Aided Design;
Daylighting; Energy Conservation; **Expert   Systems** ; HVAC Systems;
Lighting Systems; Thermal Analysis; Thermal Insulation
Identifiers: EDB/320107; EDB/990200; NTISDE
Section Headings: 89B (Building Industry Technology--Architectural Design
and Environmental Engineering); 97G (Energy--Policies, Regulations, and
Studies)

10/5/16   (Item 3 from file: 8)
DIALOG(R)File   8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

05639188   E.I. No: EIP00085292788
Title: Mobile agents and the SARA digital library
Author: Yang, Yanyan; Rana, Omer F.; Georgousopoulos, Christos; Walker,
David W.; Williams, Roy
Corporate Source: Cardiff Univ, Cardiff, UK
Conference Title: ADL 2000: IEEE Advances in Digital Libraries
Conference Location: Washington, DC, USA Conference Date:
19000522-19000524
Sponsor: IEEE Computer Society; The National Library of Medicine
E.I. Conference No.: 57197
Source: Proceedings of the Forum on Research and Technology Advances in
Digital Libraries, ADL 2000. IEEE, Piscataway, NJ, USA. p 71-77
Publication Year: 2000
ISSN: 1092-9959
Language: English
Document Type: CA; (Conference Article)   Treatment: T; (Theoretical)
Journal Announcement: 0010W1
Abstract: Remote-sensing data about the Earth's environment is being
created at an ever-increasing **rate** and distributed among heterogeneous
remote **sites** . Traditional models of distributed computing are inadequate
to support such complex applications, which generally involve a large
quantity of data. In this paper, we explore an approach based on mobile

agent **techniques** for autonomous data processing and information discovery on the Synthetic Aperture Radar Atlas (SARA) digital library, which consists of distributed multi-agency archives of multi-spectral remote-sensing imagery of the Earth. Our goal is to enable automatic and dynamic configuration of distributed parallel computing resources and to efficiently support on-demand processing of such a remote-sensing archive. The design, architecture and implementation of a prototype system that applies this approach is reported on here. (Author abstract) 18 Refs.

Descriptors: *Information services; Libraries; Synthetic aperture radar; Remote sensing; Distributed **database** systems; Parallel processing systems ; **Artificial intelligence** ; Computer systems programming; Computer architecture

Identifiers: Synthetic aperture radar atlas (SARA) digital library; Mobile agents; Multiagent systems

Classification Codes:

903.4.1 (Libraries)

903.4 (Information Services); 716.2 (Radar Systems & Equipment); 731.1 (Control Systems); 723.3 (Database Systems); 722.4 (Digital Computers & Systems)

903 (Information Science); 716 (Radar, Radio & TV Electronic Equipment) ; 731 (Automatic Control Principles); 723 (Computer Software); 722 (Computer Hardware)

90 (GENERAL ENGINEERING); 71 (ELECTRONICS & COMMUNICATIONS); 73 (CONTROL ENGINEERING); 72 (COMPUTERS & DATA PROCESSING)

10/5/17   (Item 4 from file: 8)
DIALOG(R)File   8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

04201650   E.I. No: EIP95072765085
   Title: **Equipment selection in opencast mining using an** expert   system **shell**
   Author: Haidar, Ali; Naoum, Shamil
   Corporate Source: South Bank Univ, London, Engl
   Conference Title: Proceedings of the 2nd Congress on Computing in Civil Engineering. Part 2 (of 2)
   Conference   Location: Atlanta,   GA,   USA   Conference   Date: 19950605-19950608
   Sponsor: ASCE
   E.I. Conference No.: 43201
   Source: Computing in Civil Engineering (New York) v 2 1995. ASCE, New York, NY, USA. p 1569-1576
   Publication Year: 1995
   CODEN: CCENEX
   Language: English
   Document Type: CA; (Conference Article)   Treatment: A; (Applications); T ; (Theoretical)
   Journal Announcement: 9509W1
   Abstract: The selection of the excavation and haulage equipment to remove the overburden in Opencast Mining is a major decision that involves the use of human knowledge and expertise in addition to computational **methods** . The authors' research sought to identify the variables that influence the selection of the different equipment. A theoretical model was used to assist in designing a hybrid Knowledge Based and optimization application for such a selection in order to minimize the cost of the operation. The model suggested two types of relationships: First, the type of equipment is influenced by the mine **parameters** , the soil **characteristics** , and the operating conditions of the mining site. Second, the make, number and

operating life of the selected equipment are a function of the production
rate, the ownership and capital costs as well as the equipment
characteristics . A Knowledge Base System shell for developing and
maintaining Knowledge Based and optimization applications, was considered
as the tool to solve the problem. (Author abstract) 8 Refs.
   Descriptors: ***Exper** t **systems** ; Open pit mining; Mining equipment;
Mathematical models; Knowledge based systems; Costs; Soils; Productivity;
Problem solving
   Identifiers: Opencast mining; Equipment selection; Theoretical model;
Mine **parameters** ; Soil **characteristics** ; Operating mining **site**
condition; Production **rate** ; Ownership; Capital costs
   Classification Codes:
   723.4.1 (Expert Systems)
   723.4 (Artificial Intelligence); 502.1 (Mine & Quarry Operations);
502.2 (Mine & Quarry Equipment); 483.1 (Soils & Soil Mechanics)
   723 (Computer Software); 502 (Mine & Quarry Equipment & Operations);
921 (Applied Mathematics); 911 (Industrial Economics); 483 (Soil
Mechanics & Foundations)
   72 (COMPUTERS & DATA PROCESSING); 50 (MINING ENGINEERING); 92
(ENGINEERING MATHEMATICS); 91 (ENGINEERING MANAGEMENT); 48 (ENGINEERING
GEOLOGY)

   10/5/19   (Item 6 from file: 8)
DIALOG(R)File   8:Ei Compendex(R)
(c) 2005 Elsevier Eng.   Info. Inc. All rts. reserv.

03908455   E.I. No: EIP94081354035
   Title: Site characterization by artificial neural   networks
   Author: Rizzo, Donna M.; Dougherty, David E.; Lillys, Theodore P.
   Corporate Source: Univ of Vermont, Burlington, VT, USA
   Conference Title: Proceedings of the 21st Annual Conference on Water
Policy and Management: Solving the Problems
   Conference Location: Denver, CO, USA   Conference Date: 19940523-19940526
   Sponsor: ASCE; American Water Resources Association; Center for Advanced
Decision Support for Water and Environmental Systems; Colorado State
University; Denver Water Board; et al
   E.I. Conference No.: 20568
   Source: Proceedings of the 21st Annual Conference on Water Policy and
Proc 21 Annu Conf Water Policy Manage Solving Probl 1994. Publ by ASCE, New
York, NY, USA. p 250-253
   Publication Year: 1994
   ISBN: 0-7844-0020-2
   Language: English
   Document Type: CA; (Conference Article)   Treatment: T; (Theoretical); A;
(Applications)
   Journal Announcement: 9409W3
   Abstract: Recently, an optimal groundwater management model has been
developed to treat groundwater remediation problems at Lawrence Livermore
National Laboratory (LLNL). The objective of the model is to identify the
best remediation **strategies** (well **site** selection and pumping **rates**)
so that water quality standards are met at a specified reliability level
within a given time frame. A thorough understanding of the hydrodynamic
behavior of aquifer systems requires a complete and accurate determination
of the physical **parameters** of the groundwater system. An example, the one
we will examine here, is the identification of three-dimensional hydraulic
conductivity fields for LLNL's Main Site. 9 Refs.
   Descriptors: *Groundwater; **Neural   networks** ; Management; Optimization;
Probability

Identifiers: Site characterization; Artificial **neural   networks**
Classification Codes:
  444.2   (Groundwater); 912.2   (Management); 922.1   (Probability Theory);
723.4   (Artificial Intelligence)
  444   (Water Resources); ·912   (Industrial Engineering & Management); 922
(Statistical Methods); 723   (Computer Software)
  44   (WATER & WATERWORKS ENGINEERING); 91   (ENGINEERING MANAGEMENT); 92
(ENGINEERING MATHEMATICS); 72   (COMPUTERS & DATA PROCESSING)

10/5/20   (Item 7 from file: 8)
DIALOG(R)File   8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

03593418   E.I. Monthly No: EIM9304-024187
  Title: Characteristics **of digitized images of technical articles.**
  Author: Viswanathan, Mahesh; Nagy, George
  Corporate Source: IBM/Storage Systems Products Div., San Jose, CA, USA
  Conference Title: Machine Vision Applications in Character Recognition
and Industrial Inspection
  Conference Location: San Jose, CA, USA   Conference Date: 19920210
  Sponsor: SPIE - Int Soc for Opt Engineering, Bellingham, WA, USA
  E.I. Conference No.: 17779
  Source: Proceedings of SPIE - The International Society for Optical
Engineering v 1661. Publ by Int Soc for Optical Engineering, Bellingham,
WA, USA. p 6-17
  Publication Year: 1992
  CODEN: PSISDG   ISSN: 0277-786X   ISBN: 0-8194-0815-8
  Language: English
  Document Type: PA; (Conference Paper)   Treatment: A; (Applications); T;
(Theoretical); X; (Experimental)
  Journal Announcement: 9304

  Abstract: Document image blocks characterized using projection profiles
is proposed. We have collected statistical information on scanned pages of
technical articles as a by-product of digitized document analysis.
Specifically, in our hierarchical block segmentation and labeling approach
(syntactic), 65 training and test pages from two publications were used.
Additional information on compression and profiles was also used.
Pixel-level information is required as input whether the analyzing tool is
an **expert   system** or something else. The issues covered are: (1) Profile
 characteristics of document objects like text, line drawings, tables, and
half-tones, and the variation of these profiles with block size, type size,
and direction of scan. (2) Speckle noise: sizes and distribution. (3) For
the hierarchical (syntactic) approach, number of tree nodes at each level
along with their areas, and comparison with node areas derived from
transition-cut trees. (4) CCITT-Group 4 compression statistics on document
sub-blocks and whole pages. (5) Size of postscript files and postscript
commands used in printing these page files. We believe that these results
would allow predicting some **characteristics** of a printed **page** digitized
at any specified sampling **rate** . 8 refs.
  Descriptors: *COMPUTER VISION; IMAGE ANALYSIS; IMAGE COMPRESSION;
SPURIOUS SIGNAL NOISE; STATISTICAL **METHODS** ; SPECKLE; TREES (MATHEMATICS)
  Identifiers: DOCUMENT IMAGE ANALYSIS; PROJECTION PROFILES; DIGITIZED
DOCUMENT ANALYSIS; SPECKLE NOISE; COMPRESSION STATISTICS; X-Y TREE
STATISTICS
  Classification Codes:
  723 (Computer Software); 922 (Statistical Methods); 921 (Applied
Mathematics)
  72 (COMPUTERS & DATA PROCESSING); 92 (ENGINEERING MATHEMATICS)

10/TI/1   (Item 1 from file: 2)

 Title: Extracting information from the Web for concept learning and collaborative filtering

10/TI/2   (Item 2 from file: 2)

 Title: What is a tall poppy among Web pages? [WWW information retrieval]

10/TI/3   (Item 3 from file: 2)

 Title: Proceedings of the First International Conference on The Practical Application of Knowledge Discovery and Data Mining PADD 97

10/TI/4   (Item 4 from file: 2)

 Title: Intelligent OCR editor

10/TI/5   (Item 5 from file: 2)

 Title: HiPNeT-1: a highly pipelined architecture for neural   network
 training

10/TI/6   (Item 6 from file: 2)

 Title: Landfill site selection, a microcomputer expert   system

10/TI/7   (Item 7 from file: 2)

 Title: GEOTOX: a knowledge-based system for hazardous site evaluation

10/TI/8   (Item 1 from file: 35)

DEVELOPMENT OF A COMPUTER MODEL AND EXPERT   SYSTEM **FOR PNEUMATIC FRACTURING OF GEOLOGIC FORMATIONS (PERMEABILITY)**

10/TI/9   (Item 2 from file: 35)
DIALOG(R)File 35:(c) 2005 ProQuest Info&Learning. All rts. reserv.

METHODOLOGIES FOR ESTIMATING EMISSION RATES OF HAZARDOUS GASES FROM
SINGLE POINT SOURCES (ATMOSPHERIC DISPERSION)

10/TI/10   (Item 3 from file: 35)
DIALOG(R)File 35:(c) 2005 ProQuest Info&Learning. All rts. reserv.

SPATIAL MODELING OF SUCCESSION IN A SUBTROPICAL SAVANNA (GIS)

10/TI/11   (Item 4 from file: 35)
DIALOG(R)File 35:(c) 2005 ProQuest Info&Learning. All rts. reserv.

**GEOTOX: A KNOWLEDGE-BASED SURROGATE CONSULTANT FOR EVALUATING WASTE
DISPOSAL SITES**

10/TI/12   (Item 1 from file: 6)
DIALOG(R)File   6:(c) 2005 NTIS, Intl Cpyrght All Rights Res. All rts.
reserv.

Building performance analysis using interactive multimedia concepts

10/TI/13   (Item 2 from file: 6)
DIALOG(R)File   6:(c) 2005 NTIS, Intl Cpyrght All Rights Res. All rts.
reserv.

Development of an Unstable Slope Management System
(Final rept. Jul 89-Dec 91)

10/TI/14   (Item 1 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

   Title: An   expert   system **for evaluation and remediation of rockfall**
hazards in highway cuts

10/TI/15   (Item 2 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

   Title: **New   approach to use of total coliform test for watershed**
management

10/TI/16   (Item 3 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

   Title: **Mobile agents and the SARA digital library**

10/TI/17   (Item 4 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **Equipment selection in opencast mining using an** expert system **shell**

10/TI/18   (Item 5 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **Knowledge-based system for selecting excavation groundwater control** methods

10/TI/19   (Item 6 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **Site characterization by artificial** neural   networks

10/TI/20   (Item 7 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: Characteristics **of digitized images of technical articles.**

10/TI/21   (Item 8 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **SYNAPSE - a neurocomputer that synthesizes neural algorithms on a parallel systolic engine.**

10/TI/22   (Item 9 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **GEOTOX: A KNOWLEDGE-BASED SYSTEM FOR HAZARDOUS SITE EVALUATION.**

10/TI/23   (Item 10 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **LANDFILL SITE SELECTION: A MICROCOMPUTER** EXPERT   SYSTEM .

10/TI/24   (Item 11 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **GEOTOX: A KNOWLEDGE-BASED SYSTEM FOR HAZARDOUS SITE EVALUATION.**

10/TI/25   (Item 12 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

Title: **EVALUATION OF WASTE DISPOSAL SITES USING GEOTOX.**

10/TI/26   (Item 1 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: Featuring and modelling forest recreation in Italy

10/TI/27   (Item 2 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

**Title: Recognition of splice junctions on DNA sequences by BRAIN learning
      algorithm**

10/TI/28   (Item 3 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

**Title: Sequential subtraction scintigraphy with Tc-99(m)-RBC for the early
      detection of gastrointestinal bleeding and the calculation of bleeding
      rates: Phantom and animal studies**

10/TI/29   (Item 4 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

**Title: DETERMINATION OF ELECTRON-SPIN-RESONANCE STATIC AND DYNAMIC
      PARAMETERS BY AUTOMATED FITTING OF THE SPECTRA**

10/TI/30   (Item 5 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

**Title: KINFIT-II - A NONLINEAR LEAST-SQUARES PROGRAM FOR ANALYSIS OF
      KINETIC BINDING DATA**

10/TI/31   (Item 6 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: MULTIVARIATE PROCESS ANALYSIS WITH LATTICE DATA

10/TI/32   (Item 7 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: ABSORBED MOLECULES IN MICROPOROUS HOSTS - COMPUTATIONAL ASPECTS

10/TI/33   (Item 8 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

**Title: ESTIMATING EFFECTIVE POPULATION-SIZE AND MUTATION-RATE FROM SEQUENCE
      DATA USING METROPOLIS-HASTINGS SAMPLING**

10/TI/34   (Item 9 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

**Title: INCORPORATING RULE -BASED REASONING IN THE SPATIAL MODELING OF
      SUCCESSION IN A SAVANNA LANDSCAPE**

10/TI/35   (Item 10 from file: 34)

Title: **A SHAPE-BASED AND CHEMISTRY-BASED DOCKING METHOD AND ITS USE IN
    THE DESIGN OF HIV-1 PROTEASE INHIBITORS**

10/TI/36   (Item 11 from file: 34)

Title: **IN-SITU ESTIMATION OF TRANSPORT PARAMETERS - A FIELD DEMONSTRATION**

10/TI/37   (Item 12 from file: 34)

Title: **A CORPUS-BASED STUDY OF REPAIR CUES IN SPONTANEOUS SPEECH**

10/TI/38   (Item 13 from file: 34)

Title: **A PREDICTIVE MODEL FOR AGGRESSIVE NON-HODGKINS-LYMPHOMA**

10/TI/39   (Item 14 from file: 34)

Title: **SITE SELECTION FOR NUCLEAR-PLANTS USING FUZZY DECISION-ANALYSIS**

10/3,K/1   (Item 1 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS
(c) 2005 European Patent Office. All rts. reserv.

01101616
Method for managing information on an information net
Verfahren zur Verwaltung von Information in einem Informationsnetz
Procede pour la gestion d'information dans un reseau informatique
PATENT ASSIGNEE:
    SIEMENS CORPORATE RESEARCH, INC., (1621440), 755 College Road East,
      Princeton, New Jersey 08540, (US), (Applicant designated States: all)
INVENTOR:
    Wynblatt, Michael J., 103 Endsliegh Ct., Robbinsville, NJ 08791, (US)
    Benson, Daniel C., 3356 NE 182nd, Seattle,WA 98159, (US)
LEGAL REPRESENTATIVE:
    Allen, Derek et al (55491), Siemens Shared Services Limited, c/o Siemens
      AG, P.O. Box 22 16 34, 80506 Munich, (DE)
PATENT (CC, No, Kind, Date): EP 965930 A1 991222 (Basic)
APPLICATION (CC, No, Date):   EP 99304740 990617;
PRIORITY (CC, No, Date): US 98649 980617
DESIGNATED STATES: DE; FR; GB; IT
EXTENDED DESIGNATED STATES: AL; LT; LV; MK; RO; SI
INTERNATIONAL PATENT CLASS: **G06F-017/30**
ABSTRACT WORD COUNT: 51
NOTE:
    Figure number on first page: 2

LANGUAGE (Publication,Procedural,Application): English; English; English
FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS A	(English)	199951	891
SPEC A	(English)	199951	4715
Total word count – document A			5606
Total word count – document B			0
Total word count – documents A + B			5606

INTERNATIONAL PATENT CLASS: **G06F-017/30**

...SPECIFICATION and the image with the greatest total score is chosen as
   the representative image. A **neural   net** program trained with human
   rated   pages was used to optimize the weights for a linear formula. The
    neural   net was single level, with linear weighting on the inputs,
   each input representing one of the **features** in Figure 7. Feedback from
   the training was used in the traditional iterative manner to optimize the
   weights. The **table** in Figure 7 summarizes the **features** , and gives
   some intuition as to why each **feature** is relevant in practice. Figure 8
   shows in flow chart form a suitable **process** for extracting a
   representative image in accordance with the present invention.
      Figure 9 shows some...

10/3,K/2   (Item 1 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00874877   **Image available**
METHOD AND APPARATUS FOR USER INTEREST MODELLING
PROCEDE ET DISPOSITIF DE MODELISATION DE L'INTERET D'UN UTILISATEUR

Patent Applicant/Assignee:
  SMARTHAVEN B V, Arlandaweg 92, NL-1043 EX Amsterdam, NL, NL (Residence),
    NL (Nationality), (For all designated states except: US)
Patent Applicant/Inventor:
  MASTBOOM Aiko, Groeneveen 86, NL-1103 EE Amsterdam, NL, NL (Residence),
    NL (Nationality), (Designated only for: US)
  WOLTERS Leonard Jan, Nickeriestraat 36, NL-1058 VZ Amsterdam, NL, NL
    (Residence), NL (Nationality), (Designated only for: US)
  SMITH Matthew Longshore, Kerkstraat 121-9, NL-1017 GE Amsterdam, NL, NL
    (Residence), US (Nationality), (Designated only for: US)
  KUZ Ihor Theodore, Reguliersgracht 10, NL-1017 LR Amsterdam, NL, NL
    (Residence), NL (Nationality), (Designated only for: US)
  VAN DE WIJGERD Joost, Postjeskade 125-3, NL-1058 DM Amsterdam, NL, NL
    (Residence), NL (Nationality), (Designated only for: US)
Legal Representative:
  JORRITSMA Ruurd (et al) (agent), Nederlandsch Octrooibureau,
    Scheveningseweg 82, P.O. Box 29720, NL-2502 LS The Hague, NL,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 200208989 A1 20020131 (WO 0208989)
  Application: WO 2000NL515 20000721   (PCT/WO NL0000515)
  Priority Application: WO 2000NL515 20000721
Designated States:
(Protection type is "patent" unless otherwise stated - for applications
prior to 2004)
  AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE
  ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT
  LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM
  TR TT TZ UA UG US UZ VN YU ZA ZW
  (EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE
  (OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
  (AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
  (EA) AM AZ BY KG KZ MD RU TJ TM
Publication Language: English
Filing Language: English
Fulltext Word Count: 7774

Main International Patent Class: **G06F-017/60**
Fulltext Availability:
  Claims
Claim
... f S10 3.
  The general profile is a hybrid between a Semantic Network and a **Neural
  Network** . At its most fundamental level, it contains two main objects:
  nodes and links, cf. Fig...

...for the particular skill of the PE. The PE also contains a structure
  termed an **attribute** . An **attribute** is an extension to a node,
  containing information to that node, such as past results, where the
  information was obtained from, and so on. **Attributes** that are in the PE
  can be 1 5 extensions of nodes that exist in...larger profile, with the
  appropriate logic and extra information, i.e. PE concepts and the
  attributes , to perform the skill in a highly intelligent manner. This is
  the essence of interpreting...

...connections are lost, but the information is maintained in the profile
  extension. Following the docking **procedure** is the interpretation of the
  query by this new seamless profile, a user profile, which...

...query on the user profile. A QP is a miniature UP, containing nodes,

links, and **attributes** . The idea here is to extract from the user
profile the most relevant information needed...
...be done in a variety of methods as dependent upon the current skill. One
such **technique** employed in the main embodiment of the present invention
is termed Spreading Activation. The goal...

...and links and determine the most relevant concepts pertaining to a
query. The spreading activation **technique** is a variant of Constrained
Spreading Activation.
User profiles consist of a network of interconnected...

...collection of activated nodes that are then translated into a query
profile. Note that the **attributes** that correspond to the extracted
concepts are also placed in the query proffle. Step 5...source. Obtaining
results from the query profile (Fig. 4, arrow d) is a much simpler
process . As discussed above, the query profiles contain **attributes**
that can contain several different forms of information, including
results (such as URLs, as is...

...be associated with a concept but are NOT appropriate for this particular
query. The same **methodology** can be followed to extract results from
other profiles. If a query were expanded upon...The primary embodiment of
the present invention is to endow a software agent with these
characteristics . The agent becomes personal because the user profile
(this invention) models the interests of the...

...of links to her, via the agent web browser interface. She begins to
review the **web** **pages** , and for each **page** , she provides a **rating**
(or feedback) on the quality of the **web** **page** as it pertains to her
query for a recipe. In this example, she **rates** very highly a 5 **page**
from a restaurant in Italy that lists a number of great recipes. Now the
agent...

10/3,K/3   (Item 2 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00856082
METHOD AND SYSTEM FOR SEMI-FUNGIBLE COMMODITY ITEM TRANSACTIONS
PROCEDE ET SYSTEME PERMETTANT DES TRANSACTIONS DE BIENS UTILITAIRES
    SEMI-FONGIBLES
Patent Applicant/Assignee:
  EUMEDIX COM BV, Flint, Prinsengracht 963, NL-1017 KL Amsterdam, NL, NL
    (Residence), NL (Nationality)
Inventor(s):
  LOSTIS Alain, 14, rue de Paris, F-78560 Le Port Marly, FR,
  CAPOLINO Ugo, Beethovenstraat, 4, NL-1077 JG Amsterdam, NL,
  SIDERIUS Jan, Doorpsstraat, 36, NL-3632 AT Loenen a.d. Vecht, NL,
Legal Representative:
  READ Matthew Charles (et al) (agent), Venner Shipley & Co, 20 Little
    Britain, London EC1A 7DH, GB,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 200188775 A2 20011122 (WO 0188775)
  Application: WO 2001EP5554 20010516 (PCT/WO EP0105554)
  Priority Application: US 2000573828 20000518; US 2001841020 20010424
Designated States:
(Protection type is "patent" unless otherwise stated - for applications

prior to 2004)
  AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CO CR CU CZ DE DK DM DZ
  EC EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR
  LS LT LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL
  TJ TM TR TT TZ UA UG UZ VN YU ZA ZW
  (EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
  (OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
  (AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
  (EA) AM AZ BY KG KZ MD RU TJ TM
Publication Language: English
Filing Language: English
Fulltext Word Count: 26047

Main International Patent Class: **G06F-017/60**
Fulltext Availability:
  Claims

Claim
...   understood that, if a particular supplier offers two or more producti,
  by employing a similar **methodology** the system can provide optimization
  reconimendations for different product mixes. Moreover, (inverted
  exclamation mark)t...exploitation) rights for tliree parking lots around
  its buildings. Each lot has its own unique **characteristics** and Since a
  hospital operates 24 hours/day, 7 days per week one or more...

...not use any decision making tool at all. The hospital specifies a set of
  evaluation **criteria** which identify the different needs to be taken into
  account for each of the lots...

...to have that level of service for that lot. Depending upon the
  particular implementation, these **criteria** and/or certain sub- **criteria**
  are each represented by an evaluation measure that is binary (e.g. must
  have or...

...have), linear (also called "relative")(e.g. on a scale of 0 to 100, if
  criterion Twas rated a 100 what would you rate **criterion** I-P.), or
  direct (e.g. on a scale of 0 to 100, individually rate **criteria** A
  through F. Assume that the hospital specified **criteria** are Security,
  Convenience and Parking Rate Charges, each being subject to a relative
  evaluation **criterion** . Morcover, the Security **criterion** has two sub
  criteria to be used to analyze the supplier offering: Video Camera(s)
  and 24hr/7day Guard, each being subject to a binary evaluation **criterion**
   (provided or not provided). The convenience **criterion** has three sub-
  criteria to be used to analyze the supplier offering: change machine on
  premises, live cashier, and emergency phone box in lot. Each of these is
  similarly subject to a binary evaluation **criterion** . The final
  criterion , Parking Rate Charges 3o also has two evaluation **criteria** to
  be used to analyze the supplier offering: flat fee if parked for more
  than a binary evaluation **criterion** (Le. yes or no). - 51
  As a prelude to the negotiation, the hospital is prompted...

...of Convenience and Parking Rate Charges? In response, the hospital
  provides the information shown in **Table** 24 as the "Input" value. From
  those values the `Relative Value" as a percent of the whole (in this case
  100 + 60 + 40 or 200).
   Criterion Input Relative Value
  Security 100 50%
  Binary   .

Binary
Convenience 60 30%
Binary
Binary
Binary
Parking Rate Charges 40 20%
Binary
Binary
Table 24
Within each **criterion** , sinular prompting is done for the sub- **criteria**
except lo that, in this case, the sub **criteria** for Security and Parking
Rate Charges are direct and the sub **criteria** for Convenience are
linear. At this point, the hospital decides that the Emergency Phone
criterion is no longer important. As a result, that **criterion** is
deleted. After appropriate prompting, the hospital's inputs are as shown
in **Table** 25.
Criterion to Supplier Input Relative Value
Security 100 50%
Binary 70 35%
Binary 30 15%
Convenience...

...100 25%
Binary 20 5%
Parking Rate Charges 40 20%
Binary 60 10%
Binary 10%
Table 25
The **process** is repeated for cach of the remaining two lots. - 52 Those
sub **criteria** are then used to prompt each supplier relative to their
offering. For example, Watch Dogs is prompted for cach lot and provides
the results shown in **Table** 26.
Watch Dogs **Criterion** Input Utility Input Utifity Input Utifity
Lot 1 Value Lot 2 Value Lot 3 Value...

...No
Parking Rate Charges
Binary Ye No Yes
Binary No No Yes
Total Utuity 1
Table 26
The Utility Value for each sub **criterion** is then taken as the Relative
Value for each "Yes". The total or overall service...

...value for each lot is the sum of the individual utility values as shown
in **Table** 27.
Watch·Dogs **Criterion** Input Utility Input Utility Input Utilay
Lot 1 Value Lot 2 Value Lot 3 Value...

...Binary Yes -No Yes 10
Binary No No Yes lo
Total Utility 30 45 80
Table 27
The same is done for each of the other service suppliers. Presuming that
the others have complied, **Table** 28 represents a compilation of example
results for cach of the potential exploiters.
Lot 1...

s 2302, 2304, 2306, 2308...

10/3,K/4   (Item 3 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00806392
**TECHNOLOGY SHARING DURING ASSET MANAGEMENT AND ASSET TRACKING IN A
    NETWORK-BASED SUPPLY CHAIN ENVIRONMENT AND METHOD THEREOF**
**PARTAGE TECHNOLOGIQUE LORS DE LA GESTION ET DU SUIVI DU PARC INFORMATIQUE
    DANS UN ENVIRONNEMENT DU TYPE CHAINE D'APPROVISIONNEMENT RESEAUTEE, ET
    PROCEDE ASSOCIE**
Patent Applicant/Assignee:
  ACCENTURE LLP, 1661 Page Mill Road, Palo Alto, CA 94304, US, US
    (Residence), US (Nationality)
Inventor(s):
  MIKURAK Michael G, 108 Englewood Blvd., Hamilton, NJ 08610, US,
Legal Representative:
  HICKMAN Paul L (agent), Oppenheimer Wolff & Donnelly, LLP, 38th Floor,
    2029 Century Park East, Los Angeles, CA 90067-3024, US,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 200139086 A2 20010531 (WO 0139086)
  Application: WO 2000US32310 20001122 (PCT/WO US0032310)
  Priority Application: US 99444653 19991122; US 99447623 19991122
Designated States:
(Protection type is "patent" unless otherwise stated - for applications
prior to 2004)
  AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE ES
  FI GB GE GH GM HR HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA
  MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT TZ
  UA UG UZ VN YU ZW
  (EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
  (OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
  (AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
  (EA) AM AZ BY KG KZ MD RU TJ TM
Publication Language: English
Filing Language: English
Fulltext Word Count: 156214

Main International Patent Class: **G06F-017/60**
Fulltext Availability:
  Detailed Description

Detailed Description
... comprehensive
  solutions to address these challenges. Correlation is provided by the use
  of rules
  based **inference   engines** . Event gathering and interpretation is
  typically
  performed by custom development of software interfaces which communicate
  ...

...to storing them. As discussed above, the correlation is preferably
  provided by a rules based **inference   engine** . After the events are
  correlated, a fault message is created in a fault message step...maximum
  data rate in the thousands of bits per second, and a much higher error
  rate . In fact, the combined bit rate times error rate performance of a

local cable could...

10/3,K/5   (Item 4 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00806389
**SCHEDULING AND PLANNING BEFORE AND PROACTIVE MANAGEMENT DURING MAINTENANCE
     AND SERVICE IN A NETWORK-BASED SUPPLY CHAIN ENVIRONMENT**
**PROGRAMMATION ET PLANIFICATION ANTICIPEE, ET GESTION PROACTIVE AU COURS DE
     LA MAINTENANCE ET DE L'ENTRETIEN D'UN ENVIRONNEMENT DU TYPE CHAINE
     D'APPROVISIONNEMENT RESEAUTEE**
Patent Applicant/Assignee:
  ACCENTURE LLP, 1661 Page Mill Road, Palo Alto, CA 94304, US, US
     (Residence), US (Nationality)
Inventor(s):
  MIKURAK Michael G, 108 Englewood Boulevard, Hamilton, NJ 08610, US,
Legal Representative:
  HICKMAN Paul L (agent), Oppenheimer Wolff & Donnelly, LLP, 38th Floor,
     2029 Century Park East, Los Angeles, CA 90067-3024, US,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 200139082 A2 20010531 (WO 0139082)
  Application: WO 2000US32228 20001122 (PCT/WO US0032228)
  Priority Application: US 99447625 19991122; US 99444889 19991122
Designated States:
(Protection type is "patent" unless otherwise stated - for applications
prior to 2004)
  AL AM AT AU AZ BA BB BG BR BY CA CH CN CU CZ DE DK EE ES FI GB GE GH GM
  HR HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG MK MN MW MX
  NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG UZ VN YU ZW
  (EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
  (OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
  (AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
  (EA) AM AZ BY KG KZ MD RU TJ TM
Publication Language: English
Filing Language: English
Fulltext Word Count: 152479

Main International Patent Class: **G06F-017/16**
Fulltext Availability:
  Detailed Description

Detailed Description
... assistance as detailed above in the description of a video operator.

  Self-Regulating System
  An **expert   system** monitors each call in aecordance with a preferred
  embodiment. The system includes rules that define...
...of the still connected callers inforining theni of the status change.

  Another aspect of the **expert   system** is lo ensure quality of service
  (QOS) and produce reports indicatilig both integrity and exceptions.
  Scheduling of resources is tied lo this **expert   system** , which
  regulates ' whether calls can be scheduled based on available or
  projected resources at the...

10/3,K/6   (Item 5 from file: 349)

00806384
**NETWORK AND LIFE CYCLE ASSET MANAGEMENT IN AN E-COMMERCE ENVIRONMENT AND
    METHOD THEREOF**
**GESTION D'ACTIFS DURANT LE CYCLE DE VIE ET EN RESEAU DANS UN ENVIRONNEMENT
    DE COMMERCE ELECTRONIQUE ET PROCEDE ASSOCIE**
Patent Applicant/Assignee:
  ACCENTURE LLP, 1661 Page Mill Road, Palo Alto, CA 94304, US, US
    (Residence), US (Nationality)
Inventor(s):
  MIKURAK Michael G, 108 Englewood Blvd., Hamilton, NJ 08610, US,
Legal Representative:
  HICKMAN Paul L (agent), Oppenheimer Wolff & Donnelly, LLP, 38th Floor,
    2029 Century Park East, Los Angeles, CA 90067-3024, US,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 200139030 A2 20010531 (WO 0139030)
  Application: WO 2000US32324 20001122 (PCT/WO US0032324)
  Priority Application: US 99444775 19991122; US 99447621 19991122
Designated States:
(Protection type is "patent" unless otherwise stated - for applications
prior to 2004)
  AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CU CZ DE DK DZ EE ES FI GB
  GE GH GM HR HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG MK
  MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG UZ VN
  YU ZW
  (EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
  (OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
  (AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
  (EA) AM AZ BY KG KZ MD RU TJ TM
Publication Language: English
Filing Language: English
Fulltext Word Count: 171499

Main International Patent Class: **G06F-017/60**

 10/3,K/7   (Item 6 from file: 349)

00806383
**COLLABORATIVE CAPACITY PLANNING AND REVERSE INVENTORY MANAGEMENT DURING
    DEMAND AND SUPPLY PLANNING IN A NETWORK-BASED SUPPLY CHAIN ENVIRONMENT
    AND METHOD THEREOF**
**PLANIFICATION EN COLLABORATION DES CAPACITES ET GESTION ANTICIPEE DES
    STOCKS LORS DE LA PLANIFICATION DE L'OFFRE ET DE LA DEMANDE DANS UN
    ENVIRONNEMENT DE CHAINE D'APPROVISIONNEMENT FONDEE SUR LE RESEAU ET
    PROCEDE ASSOCIE**
Patent Applicant/Assignee:
  ACCENTURE LLP, 1661 Page Mill Road, Palo Alto, CA 94304, US, US
    (Residence), US (Nationality)
Inventor(s):
  MIKURAK Michael G, 108 Englewood Blvd., Hamilton, NJ 08610, US,
Legal Representative:
  HICKMAN Paul L (agent), Oppenheimer Wolff & Donnelly, LLP, 1400 Page Mill
    Road, Palo Alto, CA 94304, US,

Patent and Priority Information (Country, Number, Date):
  Patent: WO 200139029 A2 20010531 (WO 0139029)
  Application: WO 2000US32309 20001122 (PCT/WO US0032309)
  Priority Application: US 99444655 19991122; US 99444886 19991122
Designated States:
(Protection type is "patent" unless otherwise stated - for applications
prior to 2004)
  AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE ES
  FI GB GE GH GM HR HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA
  MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT TZ
  UA UG UZ VN YU ZW
  (EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
  (OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
  (AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
  (EA) AM AZ BY KG KZ MD RU TJ TM
Publication Language: English
Filing Language: English
Fulltext Word Count: 157840

Main International Patent Class: **G06F-017/60**
Fulltext Availability:
  Detailed Description

Detailed Description
... to storing them. As discussed above, the correlation is preferably
  provided by a rules based **inference   engine** . After the events are
  correlated, a fault message is created in a fault message step...

   10/3,K/8   (Item 7 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00806382
METHOD FOR AFFORDING A MARKET SPACE INTERFACE BETWEEN A PLURALITY OF
   MANUFACTURERS AND   SERVICE PROVIDERS AND INSTALLATION MANAGEMENT VIA A
   MARKET SPACE INTERFACE
PROCEDE DE MISE A DISPOSITION D'UNE INTERFACE D'ESPACE DE MARCHE ENTRE UNE
   PLURALITE DE FABRICANTS ET DES FOURNISSEURS DE SERVICES ET GESTION
   D'UNE INSTALLATION VIA UNE INTERFACE D'ESPACE DE MARCHE
Patent Applicant/Assignee:
  ACCENTURE LLP, 1661 Page Mill Road, Palo Alto, CA 94304, US, US
    (Residence), US (Nationality)
Inventor(s):
  MIKURAK Michael G, 108 Englewood Blvd., Hamilton, NJ 08610, US,
Legal Representative:
  HICKMAN Paul L (et al) (agent), Oppenheimer Wolff & Donnelly LLP, 1400
    Page Mill Road, Palo Alto, CA 94304, US,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 200139028 A2 20010531 (WO 0139028)
  Application: WO 2000US32308 20001122 (PCT/WO US0032308)
  Priority Application: US 99444773 19991122; US 99444798 19991122
Designated States:
(Protection type is "patent" unless otherwise stated - for applications
prior to 2004)
  AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE
  ES FI GB GE GH GM HR HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV
  MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT
  TZ UA UG UZ VN YU ZW

```
(EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
(OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
(AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
(EA) AM AZ BY KG KZ MD RU TJ TM
```
Publication Language: English
Filing Language: English
Fulltext Word Count: 170977

Main International Patent Class: **G06F-017/60**
Fulltext Availability:
  Detailed Description

Detailed Description
... office. For example, U.S. Pat. No. 4,086,434 discloses a remote
  condition reporting **system** including a microprocessor with memory and a
  firmware program, telephone dialing equipment, a clock, and...4412, the
  current switch also transports the call 3602 to the next switch under
  normal **procedures** which consists of sending an IAM or setup message to
  the next switch without the NCED recorded as part of the **parameter** .
  After transporting the call 3602, the current switch proceeds to step
  4418, thereby exiting the...the important information exists only in the
  hidden relationships among items in-the databases. Recently, **artificial
  intelligence** techniques have been employed to assist users in
  discovering these relationships and, in some cases...

 10/3,K/9   (Item 8 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00545200   **Image available**
METHOD AND SYSTEM FOR DERIVING COMPUTER USERS' PERSONAL INTERESTS
PROCEDE ET SYSTEME DE DETERMINATION DES CENTRES D'INTERET DES INTERNAUTES
Patent Applicant/Assignee:
  RULESPACE INC,
Inventor(s):
  KAWASAKI Charles,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 200008573 A1 20000217 (WO 0008573)
  Application: WO 99US17654 19990804 (PCT/WO US9917654)
  Priority Application: US 9895296 19980804
Designated States:
(Protection type is "patent" unless otherwise stated – for applications
prior to 2004)
```
  AE AL AM AT AU AZ BA BB BG BR BY CA CH CN CR CU CZ DE DK EE ES FI GB GD
  GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG
  MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG UZ VN
  YU ZA ZW GH GM KE LS MW SD SL SZ UG ZW AM AZ BY KG KZ MD RU TJ TM AT BE
  CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN
  GW ML MR NE SN TD TG
```
Publication Language: English
Fulltext Word Count: 9402

Main International Patent Class: **G06F-017/30**
Fulltext Availability:
  Detailed Description

Detailed Description
... the rated lists at 124 in FIG. 8 for each of the training pages.

A **neural** – **network** 130 receives the page ratings (good or bad) via path 132 .from the lists 124 and the weighted lists 120. It also accesses the weight **database** 1 10. The **neural** – **network** then executes a series of equations for analyzing the entire set of training pages (for example, 10,000 web pages) using the set of weightings (**database** 1 1 0) which initially are set to random values. The network processes this data...

...the accuracy of the rating. This is known as a feed-forward or back-propagation **technique** , indicated at path 134 in the drawing. This type of **neural** – **network** training arrangement is known in prior art for other applications. For example, a **neural network** software packaged called "SNNS" is available on the intemet for downloading from the University of...

10/3,K/10   (Item 9 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00207472   **Image available**
IMPROVED MEMORY SYSTEM
SYSTEME DE MEMOIRE AMELIORE
Patent Applicant/Assignee:
  HYATT Gilbert P,
Inventor(s):
  HYATT Gilbert P,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 9204673 A1 19920319
  Application: WO 91US6285 19910903   (PCT/WO US9106285)
  Priority Application: US 9041 19900904
Designated States:
(Protection type is "patent" unless otherwise stated – for applications prior to 2004)
  AT BE CA CH DE DK ES FR GB GR IT JP KR LU NL SE
Publication Language: English
Fulltext Word Count: 137004

Main International Patent Class: **G06F-012/02**
Fulltext Availability:
  Detailed Description

Detailed Description
...   least portions of different database pages stored in the same block of memory; a database **page** change can imply probable or possible re-addressing. Also, in a filter system having a...about four memory refresh operations can be invoked in the 950-ns time available.

An **artificial   intelligence** memory processor memory refresh detector can be implemented to invoke memory refreshing on a time available basis in an **artificial   intelligence** processor configuration that is suitable for time available memory refreshing. For example, if the processing...

...that time. Also, if the processing operations are relatively slower than memory speed; then an **artificial   intelligence** processing refresh detector can be implemented to detect .5 the time available inbetween

processing operations to invoke refresh operations.

For example; **artificial   intelligence** processing of inference information may perform one inference operation each microsecond. However, the aboves described...cycles from filter processor, signal processor, or array processor operations to invoke refresh operations.

An **artificial   intelligence** processor memory refresh detector can be implemented to invoke memory refreshing on a cycle stealing basis in an **artificial   intelligence** processor configuration that needs cycle stealing memory refreshing. For example, if the processor operations are ...

...used; then there may not be sufficient time available for time available refreshing, Hence,, an **artificial   intelligence** processor refresh detector can be implemented to detect cycle stealing times for stealing cycles from **artificial   intelligence** processor operations to invoke refresh operations.

A stored program computer cycle stealing memory refresh detector...

12/3,K/1   (Item 1 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02517006   SUPPLIER NUMBER: 76156234   (USE FORMAT 7 OR 9 FOR FULL TEXT)
Quova's GeoPoint.(Company Business and Marketing)
Carr, Jim
Network Magazine, 32
July 1, 2001
ISSN: 1093-8001   LANGUAGE: English   RECORD TYPE: Fulltext; Abstract
WORD COUNT: 936   LINE COUNT: 00079

...   a homegrown mapping facility. Deploying GeoPoint allowed the
company to move these workers into "more **strategic** " positions, says
Transue.
        GLOBAL NETWORK OF SERVERS
        Quova's GeoPoint service is based on a...

...Derald Muniz, Quova's chief technology officer. The company uses
data-mining algorithms and proprietary **artificial   intelligence** software
to correlate its constantly updated IP data with geographic location, he
says.
        With that information, stored in a **database** the company calls its
Data Delivery Server (DDS), GeoPoint can deliver real-time, **ratings**
-based, location-specific information about **Web   site** visitors to its
customers. On a country basis, Alexander says the company's product can...

12/3,K/2   (Item 1 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

03554367   Supplier Number: 109122187 (USE FORMAT 7 FOR FULLTEXT)
**Cerberian's New Filtering and Rating Technologies Provide Most Relevant and
    Accurate Web Content Filtering.**
Business Wire, p5306
Oct 22, 2003
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 878

...   database grows organically from the surfing habits of its users
through our real-time and **artificial   intelligence** technologies. The
result, increased speed and accuracy of the content rating **process** .
Unlike other filtering services which are one layer deep, Cerberian uses a
three level **process** consisting of first, a relevant database of rated and
categorized sites and domains; second, Cerberian's Dynamic Real-Time Rating
(DRTR) technology, and finally, Cerberian's Dynamic Background Rating (DBR)
process .
        "Our **process** ensures that our users are the most protected and
safe users on the Internet benefiting...

...seamlessly to the user.
        Dynamic Real-Time Rating / Dynamic Background Rating
        The Dynamic Real-Time **Rating** (DRTR) technology retrieves non-
rated   Web   pages from their host servers to be analyzed for content.
Cerberian's DRTR technology looks at...

JMB   Date: 18-Aug-05

...context for each word, links to and from the page, and other
sophisticated page analyzing **techniques** and then responds in one of two
ways. **Web sites** that can be **rated** as adult material, pornography or
other high-liability categories are rated, categorized and immediately
blocked...

...on the user's policy. These sites are then categorized in Cerberian's
master ratings **database** insuring that every other Cerberian user
worldwide has the same, up-to-date information. Sites that fall into other
categories are moved to Cerberian's DBR **process** for additional review.
    Cerberian's Dynamic Background Rating (DBR) **process** uses a number
of proprietary ratings modules that individually rate and categorize a page
based...

12/3,K/3   (Item 2 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

03109549   Supplier Number: 82563338   (USE FORMAT 7 FOR FULLTEXT)
**Technology Catches Up to Internet Filtering Software Industry; Cerberian's
    Internet Manager Delivers Fast and Reliable Web-Based Filtering Software
    to Businesses.**
Business Wire, p0303
Feb 6, 2002
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 822

...   expensive software and hardware. Cerberian's Internet Manager is
designed to quicken the Internet filtering **process** while making it more
cost effective for companies to implement monitoring and filtering
policies."
    Cerberian...

...Unlike many of today's filtering software applications, Cerberian does
not rely exclusively on a **database** of pre-viewed and categorized URLs.
Cerberian's DRTR service, based on **Neural Net** (**artificial
intelligence**) technology, is an embedded and online service that
dynamically reads, **rates** and categorizes **Web sites** on the fly.
    If the URL entered by the user is not in the database...

...appeared in the early 1990s," Smith said.
    "Along with nearly eliminating the slow manual reviewing **process** ,
Cerberian has lowered the costs, reduced maintenance and developed a
solution for accurately reading and...

12/3,K/4   (Item 3 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

03046796   Supplier Number: 80016381   (USE FORMAT 7 FOR FULLTEXT)
RuleSpace Wins PC Magazine Technical Excellence Award.
Business Wire, p0482
Nov 13, 2001
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade

Word Count: 717

...   or workplace. Contexion Services scales successfully with the
exponential growth of the Web by employing **techniques** based on **neural
network** technology that categorize Web content continually at record
speed. Deployed by leading companies such as...

...number of sites increasing daily, it's a tough job. RuleSpace Contexion
Services uses patented **neural   network** technology to work 24/7,
categorizing content . . . Earlier this year, AOL incorporated Contexion
Services into...

...said Dan Lulich, CTO at RuleSpace and a key developer of the world's
first **neural   network** computer. "We are honored to be recognized as a
provider of revolutionary technology that is...
...categorizing Web sites - analyzing pages, site structure, subdirectories
and links on the site. The unique **features** --patterns of letters, words
and phrases--of a given Web page are analyzed in real-time by the
application of patented **neural   network** technology and pattern
recognition **techniques** to determine if the page belongs to specific
content categories. The results of the analysis are then aggregated to
infer an overall category **rating** for **Web   sites** and subdirectories and
stored in the largest **database** of pre-categorized inappropriate Web sites
and subdirectories available today.
         RuleSpace provides enterprise access control...

 12/3,K/5   (Item 4 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

03043897   Supplier Number: 79890986   (USE FORMAT 7 FOR FULLTEXT)
BudgetLife Launches Expert Quoting System.
PR Newswire, pNA
·Nov 9, 2001
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 433

         For Which They Qualify
         CANTON, Mich., Nov. 9 /PRNewswire/ --
         Interlinx, LLC today announced the BudgetLife **Expert** Quoting
System , and implemented the new software at its Web site at
http://www.budgetlife.com . The **Expert** Quoting **System** uses proprietary
technology to help consumers locate life insurance **rates** for which they
qualify.
         "Most **Web   sites** selling life insurance today simply list the
lowest rates being offered by a group of...

...their weight or family health history. We solved this problem by
programming the companies' underwriting **rules** right into the software."
         According to Mr. Burt, consumers using the old **method** would often
apply with the company showing the lowest rate, only to find out eight...

...companies to locate the lowest rates for a consumer's specified age and
gender. The **Expert** Quoting **System** then presents a short list of
questions to select the rates for which the consumer...

...the complete list of rates, but consumers who want to apply online
should use our **Expert** Quoting **System** to locate the rates for which they
qualify," Mr. Burt says. "For those who choose...

12/3,K/6   (Item 5 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

02926373   Supplier Number: 76422577   (USE FORMAT 7 FOR FULLTEXT)
**Cerberian Adopts RuleSpace Technology to Enable Next-Generation Internet
  Access Management.**
Business Wire, p0108
July 10, 2001
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 874

...   Instead of purchasing off-the-shelf solutions that are based on
ineffective first-generation filtering **methods** and require installation
and frequent upgrades, parents can simply sign up online for Cerberian's...

...a given Web page are analyzed in real-time by the application of
patent-pending **neural   network** technology and pattern recognition
techniques to determine if the page belongs to specific content
categories.
        This approach enables Contexion Services...

...and subdirectories available today. The product continually analyzes
millions of Web sites through an automated **process** by considering their
page content, page structure, site relationship and links to other known
sites.
        The results of the analysis are then aggregated to infer an overall
category **rating** for **Web   sites** and subdirectories, then stored in the
database . The combination of Contexion Services' real-time categorization
and automated retrieval of site category ratings from Contexion Services'
unmatched **database** of pre-categorized content results in the highest
possible degree of filtering precision.
        About Cerberian...

12/3,K/7   (Item 6 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

02556471   Supplier Number: 63031972   (USE FORMAT 7 FOR FULLTEXT)
Hawk Holdings Launches Next Generation Internet Search Technology.
PR Newswire, pNA
June 29, 2000
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 1230

...   the area of distributed computing, focused on highly scalable
parallel processing algorithms and reduced modeling **techniques** . Dr.
Gerasoulis, a professor of computer science at Rutgers, and his team of
computer scientists...

...Teoma's algorithms employ a combination of text, link popularity and group social network status **parameters** to **rank** the **Web** **pages** and directory links. This unique approach, coupled with its highly scalable architecture, is at the...

...software developers and engineers with advanced degrees in computer science, computer engineering, pattern recognition, and **artificial intelligence** , including several of Dr. Gerasoulis' former students. Dr. Tao Yang, a computer science professor from related infrastructure services. Teoma's cutting edge technology combined with the management, technology, and **strategic** relationships it gains through the Hawk operating platform position the company advantageously for rapid market penetration and growth.
       About Hawk Holdings
       Hawk Holdings is a **strategic** alliance between Qwest Communications International, Inc., the broadband Internet communications company, and Baxter Investments LLC...

...marketplace. The Hawk Operating Platform consists of management experience, industry knowledge and operating expertise; the **strategic** alliance with Qwest; partner companies' enabling technology synergies; **strategic** vendor and customer relationships; and investment capital. Hawk concentrates on the financial services and media...

   12/3,K/8   (Item 7 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

02232331   Supplier Number: 57569921   (USE FORMAT 7 FOR FULLTEXT)
**Mindmaker, Inc.'S DecArt 1.01M Offers Easy Addition of Decision Support To
   Applications and Web Sites.**
PR Newswire, p5176
Nov 15, 1999
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 520

...   Using the SDK, the systems administrator can easily add multi-user decision support to a **Web** **site** . Databases are stored and the **ranking** runs on the server, while the user at the client computer can use a Web browser to specify decision **parameters** , and to display the ranked decision alternatives. In this way, multiple users can simultaneously connect to a shared **database** of decision alternatives to select the best alternative.
       "DecArt 1.01M SDK is the only...

...available in January 2000 and will be licensed from Cygron and Mindmaker through systems integrators, **strategic** development and consulting partners.
       About Mindmaker Inc.
       Mindmaker, Inc. was founded in 1996, with the mission to be the leader in Intelligent Assistant-Centric Computing, providing **Artificial Intelligence** – enabled development platforms, AI-enabled applications and Intelligent Assistants for the home and business user...

10/3,K/1
DIALOG(R)File 256:TecInfoSource
(c) 2005 Info.Sources Inc. All rts. reserv.

00146118   DOCUMENT TYPE: Review

PRODUCT NAMES: Google (750026); Yahoo! (584622); Ask Jeeves (743241)

TITLE: More than Search
AUTHOR: Blumberg, Robert Atre, Shaku
SOURCE: DM Review,   v13 n3 p42(5) Mar 2003
ISSN: 1067-3717
HOMEPAGE: http://www.dmreview.com

RECORD TYPE: Review
REVIEW TYPE: Product Analysis
GRADE: Product Analysis, No Rating

REVISION DATE: 20030730

...for Web, site, and enterprise search, which are three related, but
separate, markets. Google's **method** , which **ranks** **Web** **pages** by
importance as evaluated by Web users, seems to be very accurate. Google
considers the...
...use of natural language searching; and enterprise search systems, which
usually have certain described important **features** .

10/3,K/2
DIALOG(R)File 256:TecInfoSource
(c) 2005 Info.Sources Inc. All rts. reserv.

00143329   DOCUMENT TYPE: Review

PRODUCT NAMES: Insurance (832499)

TITLE: Internet Insurance Applications Drive Down Costs
AUTHOR: De Paula, Matthew
SOURCE: Bank Technology News,   v15 n10 p24(1) Oct 2002
ISSN: 1060-3506

RECORD TYPE: Review
REVIEW TYPE: Product Analysis
GRADE: Product Analysis, No Rating

REVISION DATE: 20030430

First Penn-Pacific Life Insurance Company now provides a new online
pre-application **process** for term life insurance. The new tool is a huge
step forward for the industry...

...an account to fill out a form on a secured part of the company's **Web**
site and computes a **rate** class and premium quote that is likely to be
what is later underwritten in the...

...encompassing program called E-Life Express, which beginning in October
2002 will also include a **feature** that allows agents to check the status
of an pre-application online.

JMB   Date: 18-Aug-05

10/3,K/3
DIALOG(R)File 256:TecInfoSource
(c) 2005 Info.Sources Inc. All rts. reserv.

00131769   DOCUMENT TYPE: Review

PRODUCT NAMES: Domino (622419)

TITLE: Lotus In Space: NASA saves millions with Domino-based satellite...
AUTHOR: Napach, Bernice
SOURCE: Sm@rtPartner,   v4 n23   p28(3) Jun 11, 2001
ISSN: 1530-7742
HOMEPAGE: http://www.smartpartnermag.com

RECORD TYPE: Review
REVIEW TYPE: Product Analysis
GRADE: Product Analysis, No Rating

REVISION DATE: 20031030

...and to update engineering staff. The program creates reports and, based
on a set of **rules** , alerts engineers to satellite problems. SERS uses its
wireless component to contact engineers by pager...

...cell phone. Additionally, the application allows engineers to learn more
about problems at a secured **Web site** . Filters have been implemented to
prioritize alerts. Currently, the application is being used with six NASA
satellites, and Mobile Foundations is...

DESCRIPTORS: Aviation; CAE; Mobile Computing; Notes/Domino; **Process**
    Control; Wireless Internet

DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

017028704   **Image available**
WPI Acc No: 2005-353022/200536
XRPX Acc No: N05-288097
   **Method and system for receiver self-priced multimedia communication over
   the Internet and a member pool - composed of a** web   page **server,
   membership subsystem, communication subsystem, accounting subsystem and
   transaction subsyste**
Patent Assignee: FANG R (FANG-I); FANG K (FANG-I)
Inventor: FANG R; FANG K
Number of Countries: 002 Number of Patents: 002
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week	
TW 224269	B1	20041121	TW 2003132211	A	20031201	200536	B
US 20050119943	A1	20050602	US 2003481678	P	20031120	200537	
			US 2004831857	A	20040426		

Priority Applications (No Type Date): TW 2003132211 A 20031201; TW
   2003125324 A 20031212
Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
TW 224269	B1			G06F-017/60	
US 20050119943	A1			G06F-017/60	Provisional application US 2003481678

Abstract (Basic): TW 224269 B1
      NOVELTY - The invention discloses a method and system for receiver
   self-priced multimedia communication over the Internet and a member
   pool. The system is a main system containing **database** server, which
   is composed of a web server, membership subsystem, communication
   subsystem, accounting subsystem and transaction subsystem, and proceeds
   multimedia communication services over the internet and a member pool
   by bundling the main system to provide the downloaded software for
   membership end. Firstly let subscriber sign in the membership subsystem
   to become the member and acquire the exclusive membership ID. The
   member can select the only receiver's condition of the three types of
   payment flows, the collect-to-connect, pay-to-connect and free connect
   and set up the accepted **rate** for the member. No matter whether
   already proceed the stored value for payment or not, any member can
   register the communication subsystem to stand-by through using the
   member end software via internet and **page** other specific member to
   execute the multimedia communication over internet. After receiving the
   call, the communication subsystem will search the condition of the
   called member and the account balance of the party to pay in the
   database of the main system so as to set up the communication time
   span and pass the call. After the called member is already in standby
   state and accepts the call of the calling member, both parties can
   proceed multimedia communication through the bridging of the
   communication subsystem. Once the communication time is expired or
   anyone of the parties disconnects, the communication subsystem then
   forwards the called record of the said communication to the accounting
   subsystem to proceed the payment and accounting calculation based on
   the formula set up in the main system and executes the payment amount
   transfer and update of the **database** of the main system. When the
   accounting subsystem proceeds payment amount transfer, then partial
   amount based on the formula set up in the main system is kept as system
   profit. Such system and method are a state-of-art design capable of
   applying multimedia communication **technique** to convert the content

value and perceived value into the financial benefit of the members and
the system provider, delivering a commercial method enhancing
communication industry.
        DwgNo 1/1
Title Terms: METHOD; SYSTEM; RECEIVE; SELF; PRICE; COMMUNICATE; MEMBER;
    POOL; COMPOSE; WEB; **PAGE** ; SERVE; MEMBER; SUBSYSTEM; COMMUNICATE;
    SUBSYSTEM; ACCOUNT; SUBSYSTEM; TRANSACTION .
Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
File Segment: EPI

12/5/3   (Item 3 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

016807029
WPI Acc No: 2005-131310/200514
Related WPI Acc No: 2002-454734
XRAM Acc No: C05-043263
XRPX Acc No: N05-112532
    **System for guiding selection of treatment regimen comprises computing
    device with knowledge bases having information about DNA** site
    methylation and rules **for evaluating/selecting disease based on it and
    generating** ranked **listing of diseases**
Patent Assignee: EPIGENOMICS AG (EPIG-N)
Inventor: BERLIN K; OLEK A; PIEPENBROCK C
Number of Countries: 001 Number of Patents: 001
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week	
US 20050021240	A1	20050127	US 2000705302	A	20001102	200514	B
			US 2004774052	A	20040206		
			US 2004857105	A	20040528		

Priority Applications (No Type Date): US 2004857105 A 20040528; US
    2000705302 A 20001102; US 2004774052 A 20040206
Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
US 20050021240	A1		29	G06F-017/60	CIP of application US 2000705302
					CIP of application US 2004774052

Abstract (Basic): US 20050021240 A1
        NOVELTY - A system comprises a computing device having a knowledge
    base with information about methylation of selected DNA **sites** in
    cells with a known disease or medical condition and/or healthy cells
    and a second knowledge base with **rules** for evaluating and selecting
    disease based on above. The information is provided to the device and a
    ranked listing of diseases or medical conditions is generated based
    on information of knowledge bases.
        DETAILED DESCRIPTION - An INDEPENDENT CLAIM is included for a
    computer program product for guiding the selection.
        USE - The system is useful for guiding the selection of therapeutic
    treatment regimen for a disease or a medical condition (claimed) e.g.
    cancer, viral and/or bacterial infection.
        ADVANTAGE - The system allows a precise diagnosis of the disease
    that is fast and efficient, since time plays an important role in the
    survival **rate** . The diagnosis is easy to access and does not involve a
    time consuming **procedure** leading to reduced cost since unnecessary

and ineffective medication is avoided. The system can be configured to
prevent a user from receiving recommendations on new therapy options
when crucial data on the patient has not been entered.
      pp; 29 DwgNo 0/3
Title Terms: SYSTEM; GUIDE; SELECT; TREAT; REGIMEN; COMPRISE; COMPUTATION;
  DEVICE; BASE; INFORMATION; DNA; **SITE** ; METHYLATION; **RULE** ; EVALUATE;
  SELECT; DISEASE; BASED; GENERATE; **RANK** ; LIST; DISEASE
Derwent Class: B04; D16; S05; T01
International Patent Class (Main): **G06F-017/60**
International Patent Class (Additional): C12Q-001/68; G01N-033/48;
  G01N-033/50; G06F-019/00
File Segment: CPI; EPI

   12/5/4   (Item 4 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

016444973   **Image available**
WPI Acc No: 2004-602889/200458
XRPX Acc No: N04-476849
   **Double-entry bookkeeping implementing method for internet based
   accounting system, involves creating dummy record set for temporarily
   balancing record set group for protecting** database **files from recording
   single entry**
Patent Assignee: LEE H M (LEEH-I)
Inventor: LEE H M
Number of Countries: 001 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
US 20040148233 A1   20040729 US 2003686314   A   20031015 200458 B

Priority Applications (No Type Date): HK 2003100701 A 20030128
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
US 20040148233 A1   32 G06F-017/60

Abstract (Basic): US 20040148233 A1
      NOVELTY - Depending on **features** of working **pages** , one of record
   set among debit or credit record set is created as subject `S' and
   other as object `O'. Additional transaction data in the same record set
   is updated and new record set `O' is created to constitute double entry
   journal. A dummy record set for temporarily balancing the record set
   groups is created for protecting **database** files from recording single
   entry.
      DETAILED DESCRIPTION - A user interface having data capturer and
   processor which composes `AR, AP, Rec, Pay, TT, JJ' working **pages** is
   created for performing or capturing data transactions. Record set
   groups with at least two record sets one for recording credit data and
   another for recording debit data are created automatically after
   capturing data. One of record sets is created as subject `S' and other
   as object `O', each of the record sets group comprise only one subject
   and many number of objects. According to the specific **procedures** , the
   amount is posted into debit field or posted into the credit field of
   each record set. Any additional transaction data in the same set is
   simultaneously updated and new record set `O' is created to constitute
   double entry journal. A dummy record set is created for temporarily
   balancing the record set groups for protection of **database** files from

recording single entry or disordered journal entries. A new record set
`E' is created for recording data resulting from exchange difference
derived between the adoption of book **rate** and transaction **rate**
according to account types involving the record sets `S,O'. Then, the
double entry journal is converted into voucher form, which is fad back
to the general purpose computer on real time basis. A double entry
journal consisting of record sets group is stored. The record sets
comprise identification fields of user's identity and user's business
unit, account codes, input amount, currency, the converted amount in
local currency. The record sets groups is stored and updated in the
from on double entry journal into the designated **database** file in one
stoke immediately after each processing. The identification fields are
recognized for the user's access right and limitations set by the
program on the usage of **database** file. Financial reports are sent by
analyzing each of journal records on reception of reporting command and
user's identity.
       USE - For implementing double-entry bookkeeping for internet based
accounting system for multinational companies.
       ADVANTAGE - The dummy record set is automatically deleted by the
program to ensure all processed data stored into the **database** are
double entry journals while complying with the
total-debit-equal-total-credit **rule** .
       DESCRIPTION OF DRAWING(S) - DESCRIPTION OF DRAWING - The figure
shows the block diagram explaining mechanism of bookkeeping and
accounting information system working in parallel to online web based
communications.
       pp; 32 DwgNo 1/23
Title Terms: DOUBLE; ENTER; IMPLEMENT; METHOD; BASED; ACCOUNT; SYSTEM;
   DUMMY; RECORD; SET; TEMPORARY; BALANCE; RECORD; SET; GROUP; PROTECT;
   DATABASE ; FILE; RECORD; SINGLE; ENTER
Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
File Segment: EPI

   12/5/5   (Item 5 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

016261123   **Image available**
WPI Acc No: 2004-419017/200439
XRAM Acc No: C04-157273
XRPX Acc No: N04-332596
   **Managing of information privacy for enterprise by identifying application
   information describing software applic**ation, storing information in
   database **, and identifying types of information contained in or used by
   the application**
Patent Assignee: BORGIA E (BORG-I); BRESLIN J (BRES-I); DE GOTTAL G
   (DGOT-I)
Inventor: BORGIA E; BRESLIN J; DE GOTTAL G
Number of Countries: 001 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
US 20040098285 A1   20040520 US 2002411370   P   20020917 200439 B
                               US 2003664530   A   20030917

Priority Applications (No Type Date): US 2002411370 P 20020917; US
   2003664530 A 20030917

Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
US 20040098285 A1   22 G06F-017/60   Provisional application US 2002411370

Abstract (Basic): US 20040098285 A1
        NOVELTY - An information privacy for an enterprise is managed by
    identifying application information that describes a software
    application; storing the information in a **database** ; identifying types
    of information that are contained in or used by the application;
    storing the types of information in the **database** ; determining and
    storing jurisdiction information.
        DETAILED DESCRIPTION - Managing of information privacy for an
    enterprise involves identifying application information that describes
    a software application used by an enterprise; storing the application
    information in a **database** ; identifying types of information that are
    contained in or used by the application; storing the types of
    information in the **database** ; determining jurisdiction information
    that describes the jurisdictions in which the application operates;
    storing the jurisdiction information in the **database** ; identifying the
    procedures used to protect the privacy of the types of information;
    storing procedural information related to the **procedures** in the
    database ; automatically determining a compliance **rating** associated
    with the application; storing the compliance **rating** in the **database**
    ; and providing status data from the **database** . The status data
    comprises a compliance **rating** .
        An INDEPENDENT CLAIM is also included for a system for an
    enterprise to manage privacy of information comprising a user interface
    for interfacing with users of the system; a **database** server and an
    application server coupled to the user interface; and a **database** and
    an application respectively coupled to the **database** server and the
    application server.
        USE - For managing privacy of information or data.
        ADVANTAGE - The method is capable to manage and monitor the
    protection of employees' and customer's private data. It enhances
    current processes to provide a decision engine around the key data
    privacy issues providing the capability for enhanced monitoring and
    management around the risk management function.
        DESCRIPTION OF DRAWING(S) - The figure is a labeled schematic
    diagram of the method.
        pp; 22 DwgNo 1/12
Title Terms: MANAGE; INFORMATION; PRIVATE; IDENTIFY; APPLY; INFORMATION;
  DESCRIBE; SOFTWARE; APPLY; STORAGE; INFORMATION; **DATABASE** ; IDENTIFY;
  TYPE; INFORMATION; CONTAIN; APPLY
Derwent Class: B04; D16; S05; T01; T05; W01
International Patent Class (Main): **G06F-017/60**
File Segment: CPI; EPI

   12/5/6   (Item 6 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

015581009   **Image available**
WPI Acc No: 2003-643166/200361
  **Electronic commerce method using human interface navigation and client
  customized decision making** technique **and recording medium capable of
  being read by computer having program for implementing the same**
Patent Assignee: LEE N K (LEEN-I)

Inventor: LEE N K
Number of Countries: 001 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
KR 2003040940 A   20030523 KR 200171615   A   20011117 200361 B

Priority Applications (No Type Date): KR 200171615 A 20011117
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
KR 2003040940 A   1 G06F-017/60

Abstract (Basic): KR 2003040940 A
        NOVELTY - An electronic commerce method and a recording medium are
    provided to adopt an intuitive and sensitive commodity selection
    navigation based on an age of a purchaser and a purchase theme in
    accordance with the age when the purchaser selects a category of a
    commodity.
        DETAILED DESCRIPTION - A purchaser connects to a web server of a
    businessman system through a communication network(S10). The web server
    performs authentication(S20,S30). If the purchaser is a normal
    member(S40), an application server designates a right to access various
    kinds of web interfaces related to an electronic commerce(S50). The
    businessman system judges whether the purchaser selects the first
    commodity navigation or the second commodity navigation for selecting a
    category of a wanted commodity on a main **web   page** (S70,S80). If the
    purchaser selects the first commodity navigation, the purchaser selects
    a classification of a purchase-objected commodity(S90). Weight value
    scores and the final classification code of a selected commodity are
    transmitted to the application server(S120). A **ranking** is designated
    to a commodity in accordance with the purchase decision **criterion**
    (S130). A commodity list **web   page** is supplied by reflecting the
    purchase decision **criterion** of the purchaser(S140). If the purchaser
    does not input a purchase **criterion** again(S150), commodity order
    information is supplied(S160) and a price is paid(S170).
        pp; 1 DwgNo 1/10
Title Terms: ELECTRONIC; METHOD; HUMAN; INTERFACE; NAVIGATION; CLIENT;
  CUSTOMISATION; DECIDE; **TECHNIQUE** ; RECORD; MEDIUM; CAPABLE; READ;
  COMPUTER; PROGRAM; IMPLEMENT
Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
File Segment: EPI

  12/5/7   (Item 7 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

014769125   **Image available**
WPI Acc No: 2002-589829/200263
XRPX Acc No: N02-468058
  **Internet-based goods and services marketing system establishes telephone
  connection between customer and salesperson at vendor system during
  provision of goods and service information**
Patent Assignee: KLEIN T J (KLEI-I); SPETNER J (SPET-I)
Inventor: KLEIN T J; SPETNER J
Number of Countries: 001 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week

US 20020077924 A1 20020620 US 2000742862   A   20001220 200263 B

Priority Applications (No Type Date): US 2000742862 A 20001220
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
US 20020077924 A1   11 G06F-017/60

Abstract (Basic): US 20020077924 A1
        NOVELTY - A vendor computer system provides multiple screens
    displaying information about products and services offered by the
    vendor for sale and phone number for contacting a salesperson at the
    vendor system, to a customer computer. A telephone connection is
    established between customer and salesperson during the provision of
    the screens. A software program controls the screens after the
    telephone connection establishment.
        DETAILED DESCRIPTION - An INDEPENDENT CLAIM is included for
    Internet-based goods and services marketing method.
        USE - For marketing products such as vehicle, clothing,
    real-estate, mail order and electronic and computer equipment, and
    services such as legal advice, insurance **rate** comparison and
    financial service to customer through Internet.
        ADVANTAGE - The system has simple construction and design and is
    easily used with highly reliable results. By allowing the customer and
    a salesperson at the vendor computer system to communicate, the
    salesperson can easily discuss the product and service **features** to
    the customer and thus can sell products through Internet easily. The
    customer who is not familiar with the **web** **site** can also acquire the
    required information for navigating the **web** **sites** from the
    salesperson.
        DESCRIPTION OF DRAWING(S) - The figure shows the flowchart
    illustrating the Internet-based goods and services marketing **procedure**

        pp; 11 DwgNo 7/8
Title Terms: BASED; GOODS; SERVICE; MARKET; SYSTEM; ESTABLISH; TELEPHONE;
  CONNECT; CUSTOMER; VENDING; SYSTEM; PROVISION; GOODS; SERVICE;
  INFORMATION
Derwent Class: T01; T05; W01; W05
International Patent Class (Main): **G06F-017/60**
File Segment: EPI

   12/5/8   (Item 8 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

014697519
WPI Acc No: 2002-518223/200255
XRPX Acc No: N02-410112
   Internet contextual search engine-based sales commission system, has
   performance-based tracking and reporting system that identifies actual
   sales, and provides credit to the affiliate
Patent Assignee: HYDER A D (HYDE-I)
Inventor: HYDER A D
Number of Countries: 001 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
US 20020055894 A1   20020509 US 2000185615   A   20000229 200255 B
                              US 2001796315   A   20010228

Priority Applications (No Type Date): US 2000185615 P 20000229; US
  2001796315 A 20010228
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
US 20020055894 A1   9 G06F-017/60   Provisional application US 2000185615

Abstract (Basic): US 20020055894 A1
        NOVELTY - A global search engine based affiliate network presents
    users of at least one affiliate with a dialog box with preloaded text
    . that is relevant to the content of a corresponding **web site** . A
    performance-based tracking and reporting system identifies actual
    sales, and provides credit to the affiliate that referred a buying user
    to the retailer.
        DETAILED DESCRIPTION - An INDEPENDENT CLAIM is also included for a
    method of generating qualified buyers for a retail **web site** .
        USE - Internet contextual search engine-based sales commission
    system.
        ADVANTAGE - Uses artificial intelligence to conduct dialog with the
    user. Utilizes proprietary **matrix** algorithm to determine the number
    of and type of searches needed to find the answer. Utilizes proprietary
    methodology to **prioritize** the search results based on the tightness
    of fit not location or frequency or popularity of keywords.
        pp; 9 DwgNo 0/0
Title Terms: SEARCH; ENGINE; BASED; SALE; COMMISSION; SYSTEM; PERFORMANCE;
  BASED; TRACK; REPORT; SYSTEM; IDENTIFY; ACTUAL; SALE; CREDIT
Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
International Patent Class (Additional): G06F-015/173
File Segment: EPI

   12/5/9   (Item 9 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

014695318   **Image available**
WPI Acc No: 2002-516022/200255
    **Method for implementing internet based noise measurement and noise
    environment** database **system**
Patent Assignee: LEE S G (LEES-I); SHIN I H (SHIN-I)
Inventor: LEE S G; SHIN I H
Number of Countries: 001 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
KR 2002005803 A   20020118 KR 200039219   A   20000710 200255 B

Priority Applications (No Type Date): KR 200039219 A 20000710
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
KR 2002005803 A   1 G06F-017/60

Abstract (Basic): KR 2002005803 A
        NOVELTY - A method for implementing Internet based noise
    measurement and constructing a noise environment **database** system are
    provided to supply a **technique** for measuring a noise **rate** by a
    micro phone and a PC by supplying a **technique** for measuring a noise
    of an external environment of user using the PC, a general mike, and a
    program on the Internet.
        DETAILED DESCRIPTION - User information as geographic information

and an external environment by connecting a user to the corresponding
web   site is inputted, and user system environment information of
hardware information as a kind of a sound card and a microphone is
inputted. A noise **rate** measuring program based on the Internet is
downloaded through the Internet in a server computer and a noise
measured using a micro phone being connected to a user computer is
received as an analogue voltage signal, and a digital signal of a
measured noise being transmitted from the sound card is read(3). A
measured noise value is corrected for correcting a distortion of a
noise by the micro phone and the sound card being connected to the user
computer(4). A necessary noise **rate** is calculated using the digital
value received from the sound card(5). The user computer displays a
noise **rate** of an external environment of the user(6) and transmits
the noise **rate** to a noise measuring and a noise environment **database**
 of the server computer through an Internet line(7). The noise **rate**
is classified in accordance with information inputted by the user(8)
and the classified noise **rate** is stored in the **database** (9).
        pp; 1 DwgNo 1/10
Title Terms: METHOD; IMPLEMENT; BASED; NOISE; MEASURE; NOISE; ENVIRONMENT;
  DATABASE ; SYSTEM
Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
File Segment: EPI

   12/5/10   (Item 10 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

013949416
WPI Acc No: 2001-433630/200147
XRAM Acc No: C01-131281
XRPX Acc No: N01-321310
   Assessment of level of contamination in soil and water involves using
   matrix of risk factors together with sensitivity and vulnerability
   aspects to determine corrective measures needed for effective
   decontamination
Patent Assignee: JENA-GEOS ING GMBH (JENA-N)
Inventor: GROSSMANN J; GRUNEWALD V; SCHAUBS A; WEIHRAUCH F
Number of Countries: 001 Number of Patents: 002
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
DE 10004065   A1   20010712   DE 1004065   A   20000131   200147   B
DE 20022760   U1   20020328   DE 1004065   A   20000131   200229
                               DE 2000U2022760 U   20000131

Priority Applications (No Type Date): DE 1062721 A 19991223
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
DE 10004065 A1   8 B09C-001/00
DE 20022760 U1   B09C-001/00   Application no.   DE 1004065

Abstract (Basic): DE 10004065 A1
     NOVELTY - To determine the comparative measures to be taken in
   decontamination of soil and water, a **matrix** is drawn up showing the
   risk factors and vulnerability. The risk categories are arranged on a
   diagonal, giving the main diagonals for a medium risk level. The
   graduations of the risks are related to protection, and the graduations

of the vulnerability are related to application.

DETAILED DESCRIPTION - A classified risk **rating** is given for each contamination by its position in the **matrix** according to the risk factors. The comparisons are shown to avoid reductions in high and very high risk factors and lowering by more than one risk stage, and alterations within the low-risk ranges.

Preferred **Features** : The risk factors are graduated into three stages in relation to effects related to protection, and sensitivity related to use. The graduation of the risk factors, for the protection of the soil and/or water is related to the relevant soil functions on the **site** and the protective materials for soil protection, and the water relevance for surface and ground water conditions. The decontamination of water takes the risk of the spread of hazardous matter and the sensitivity or vulnerability of the water resource value.

To determine the tendency of the hazardous matter to spread, a mean determination of the reactive geogene complexes is taken by measurement of concentrations along the water flow direction of the contaminated water surface at least at two points in the ground water at a given point in time, to give the level of concentration value of ic according to the expression (i) together with the unit (ii) or without dimensions on standardizing the distance at 100 km.

$$ic=(c2/c1)/(s2-s1) \quad (i)$$
$$ic=((mg/l/mg/l)/m)=(1/m)$$

s1=distance between the emission center and the first measurement point,

s2=distance between the emission center and the second measurement point,

c1=measured concentration at the first measurement point,

c2=measured concentration at the second measurement point.

The process is repeated with two further measurements at a different point in time, and a further concentration level value is calculated. An objective assessment is made of the risk according to the spread of hazardous matter by space and time, and the resource value of the water, to form a decision **matrix** for the relevant risk categories. At least the second measurement point is on a theoretical line from the contaminated zone through the first measurement point, exactly on the water flow direction line.

USE - The **technique** is to establish the level of contamination and the risk factors, to be used for decontamination of affected soil and water.

ADVANTAGE - The method takes all the factors into account to establish the concentration and nature of the contamination, for corrective measures to be taken to meet health and safety regulations.

pp; 8 DwgNo 0/3

Title Terms: ASSESS; LEVEL; CONTAMINATE; SOIL; WATER; **MATRIX** ; RISK; FACTOR; SENSITIVE; VULNERABLE; **ASPECT** ; DETERMINE; CORRECT; MEASURE; NEED; EFFECT; DECONTAMINATE

Derwent Class: D15; P35; P43; Q42; T01

International Patent Class (Main): B09C-001/00

International Patent Class (Additional): A62D-003/00; C02F-001/00; E02B-003/00; **G06F-017/60**

File Segment: CPI; EPI; EngPI

12/5/11   (Item 11 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

013808790   **Image available**
WPI Acc No: 2001-293002/200131
Related WPI Acc No: 2001-292999
XRPX Acc No: N01-209532
   Rapid procedure **for insuring against risks involved in purchasing**
   online in which a potential transaction is checked on an insurance
   database **of companies registered with the insurance provider to check**
   credit- rating **etc.**
Patent Assignee: IMPACT BUSINESS & TECHNOLOGY CONSULTING (IMPA-N)
Inventor: HAFENBRADL U; NOEL J
Number of Countries: 025 Number of Patents: 001
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week
EP 1093071	A1	20010418	EP 99120292	A	19991012	200131 B

Priority Applications (No Type Date): EP 99120292 A 19991012
Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
EP 1093071	A1	G	6	G06F-017/60	

   Designated States (Regional): AL AT BE CH CY DE DK ES FI FR GB GR IE IT
   LI LT LU LV MC MK NL PT RO SE SI

Abstract (Basic): EP 1093071 A1
       NOVELTY - **Procedure** in which web-based e-commerce transactions
   are insured against by a very rapid additional insuring step where
   details of the proposed transaction are sent from the potential
   purchaser to an insurance databank on a server computer or network and
   a search is carried out for the sellers details in the databank. If the
   necessary conditions are fulfilled then the sale is authorized.
       DETAILED DESCRIPTION - When a purchaser wishes to buy something
   from a sellers **web - site** details relevant to the purchase (4) are
   sent to a databank (3). Here a search is carried out in the databank
   for details relevant to the seller such as his credit **rating** . An
   insurance or policy number (5) is granted if the transaction is
   successful. When the check is unsuccessful no number is granted and an
   appropriate message is sent over the data network (World Wide Web using
   HTML). A company wishing to sell over the Internet pays a premium to an
   insurance company that then registers its details and sets a
   transaction credit limit for it in a **database** .
       USE - Granting of insurance to cover the risk of making online
   purchases over the Internet.
       ADVANTAGE - The insurance **procedure** is very quick taking
   optimally between 1 and 5 seconds. Internet e-commerce is encouraged as
   risks for the buyer are reduced.
       DESCRIPTION OF DRAWING(S) - Figure shows a schematic view of the
   arrangement.
       purchaser PC (1)
       seller's **homepage** (6)
       World Wide Web (2)
       insurance databank computer (3)
       proposed purchase details (4)
       insurance number (5)
       E-mail server for notifying seller. (7)
       pp; 6 DwgNo 1/1
Title Terms: RAPID; **PROCEDURE** ; ENSURE; RISK; PURCHASE; POTENTIAL;
   TRANSACTION; CHECK; INSURANCE; **DATABASE** ; COMPANY; REGISTER; INSURANCE;
   CHECK; CREDIT; **RATING**
Derwent Class: T01

International Patent Class (Main): **G06F-017/60**
File Segment: EPI

12/5/12   (Item 12 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

013808787   **Image available**
WPI Acc No: 2001-292999/200131
XRPX Acc No: N01-209529
   Rapid procedure **for insuring against risks involved in purchasing
   online in which a potential transaction is checked on an insurance**
   database **of companies registered with the insurance provider to check
   credit-** rating **etc.**
Patent Assignee: IMPACT BUSINESS & TECHNOLOGY CONSULTING (IMPA-N)
Inventor: HAFENBRADL U; NOEL J
Number of Countries: 025 Number of Patents: 001
Patent Family:
Patent No   Kind   Date   Applicat No   Kind   Date   Week
EP 1093066   A1   20010418 EP 2000121824   A   20001006 200131 B

Priority Applications (No Type Date): EP 99120292 A 19991012
Patent Details:
Patent No Kind Lan Pg   Main IPC   Filing Notes
EP 1093066   A1 G   7 G06F-017/60
   Designated States (Regional): AL AT BE CH CY DE DK ES FI FR GB GR IE IT
   LI LT LU LV MC MK NL PT RO SE SI

Abstract (Basic): EP 1093066 A1
       NOVELTY - **Procedure** in which web-based e-commerce transactions
   are insured against by a very rapid additional insuring step where
   details of the proposed transaction are sent from the potential
   purchaser to an insurance databank on a server computer or network and
   a search is carried out for the sellers details in the databank. If the
   necessary conditions are fulfilled then the sale is authorized.
       DETAILED DESCRIPTION - When a purchaser wishes to buy something
   from a sellers **web - site** details relevant to the purchase (4) are
   sent to a databank (3). Here a search is carried out in the databank
   for details relevant to the seller such as his credit **rating** . An
   insurance or policy number (5) is granted if the transaction is
   successful. When the check is unsuccesful no number is granted and an
   appropriate message is sent over the data network (World Wide Web using
   HTML). A company wishing to sell over the Internet pays a premium to an
   insurance company that then registers its details and sets a
   transaction credit limit for it in a **database** .
       USE - Granting of insurance to cover the risk of making online
   purchases over the Internet.
       ADVANTAGE - The insurance **procedure** is very quick taking
   optimally between 1 and 5 seconds. Internet e-commerce is encouraged as
   risks for the buyer are reduced.
       DESCRIPTION OF DRAWING(S) - Figure shows a schematic view of the
   arrangement.
       purchaser PC (1)
       seller's **homepage** (6)
       World Wide Web (2)
       insurance databank computer (3)
       proposed purchase details (4)
       insurance number (5)

E-mail server for notifying seller. (7)
pp; 7 DwgNo 1/2
Title Terms: RAPID; **PROCEDURE** ; ENSURE; RISK; PURCHASE; POTENTIAL;
  TRANSACTION; CHECK; INSURANCE; **DATABASE** ; COMPANY; REGISTER; INSURANCE;
  CHECK; CREDIT; **RATING**
Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
File Segment: EPI

12/5/13   (Item 13 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

013284303   **Image available**
WPI Acc No: 2000-456238/200040
XRPX Acc No: N00-340262
  Display advertising selection procedure **for use in world wide web,
  involves selecting advertisement to display based on the display
  probability of each advertisement opposing to each estimated** attribute
Patent Assignee: NEC CORP (NIDE)
Inventor: ABE N; NAKAMURA A
Number of Countries: 002   Number of Patents: 003
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week	
JP 2000163477	A	20000616	JP 98337649	A	19981127	200040	B
JP 3389948	B2	20030324	JP 98337649	A	19981127	200323	
US 6591248	B1	20030708	US 99447946	A	19991129	200353	

Priority Applications (No Type Date): JP 98337649 A 19981127
Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
JP 2000163477	A		13	G06F-017/60	
JP 3389948	B2		13	G06F-017/60	Previous Publ. patent JP 2000163477
US 6591248	B1			G06F-017/60	

Abstract (Basic): JP 2000163477 A
     NOVELTY - The advertising selection unit (151) has the
characteristics of deforming the objective function maximization
problem with restrictions which is deformed into the format of
transportation problem. The maximization unit applies the solution of
transportation problem and it selects the advertisement to display
based on the display probability of each advertisement opposing to each
estimated **attribute** .
     DETAILED DESCRIPTION - The input probability of each **attribute**
and the **rate** of click of each advertisement opposing to each
attribute are estimated from click performance.
     USE - For selecting advertisement suitable for displaying to **web
page** .
     ADVANTAGE - Since the advertisement is chosen based on the display
probability, the entire frequency of the click is increased.
     DESCRIPTION OF DRAWING(S) - The figure shows the block diagram of
display advertising selection system.
     Advertising selection unit (151)
     pp; 13 DwgNo 1/8
Title Terms: DISPLAY; ADVERTISE; SELECT; **PROCEDURE** ; WORLD; WIDE; WEB;
  SELECT; ADVERTISE; DISPLAY; BASED; DISPLAY; PROBABILITY; ADVERTISE;
  OPPOSED; ESTIMATE; **ATTRIBUTE**

Derwent Class: T01
International Patent Class (Main): **G06F-017/60**
International Patent Class (Additional): G06F-013/00; G06F-019/00
File Segment: EPI

12/5/14   (Item 1 from file: 347)
DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

07240772   **Image available**
METHOD FOR MEDIATING PRIVATELY-SUBSCRIBED BOND OF SMALL NUMBER PEOPLE

PUB. NO.:   2002-109223 [JP 2002109223 A]
PUBLISHED: April 12, 2002 (20020412)
INVENTOR(s):   TAKAHASHI YOSHIAKI
APPLICANT(s): TAKAHASHI YOSHIAKI
APPL. NO.:   2000-293520 [JP 2000293520]
FILED: September 27, 2000 (20000927)
INTL CLASS: **G06F-017/60**

ABSTRACT

PROBLEM   TO   BE   SOLVED: To   provide   a   method   for   mediating   a
privately-subscribed bond of a small number people which is issued by a
small business through the Internet.

SOLUTION: On a **homepage** operated by a mediator, a market for collecting
the privately-subscribed bond of a small number people is divided into a
general market 11 where a corporate bond issuing company presents a
corporate bond issuing condition and a special market 12 where a corporate
bond purchase desiring person presents the **rate** of interest. In the
special market 12, in particular, the corporate bond issuing condition
except the **rate** of interest of each corporate bond issuing company and
financial statements for each company are made into a **database** together
with the name of each company so as to be freely read by the corporate bond
purchase desiring person. A corporate bond issuing **procedure** is started
on condition that the desiring person U presents the **rate** of interest and
the issuing company K performs approval.

12/5/15   (Item 2 from file: 347)
DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

07223559   **Image available**
HIGHEST- **RANK** INFORMATION RETRIEVING SYSTEM

PUB. NO.:   2002-091998 [JP 2002091998 A]
PUBLISHED: March 29, 2002 (20020329)
INVENTOR(s):   KONISHI YUTAKA
               SENDA YASUHIRO
APPLICANT(s): WORLD ECONOMIC INFORMATION SERVICES
APPL. NO.:   2000-285457 [JP 2000285457]
FILED: September 20, 2000 (20000920)
INTL CLASS: G06F-017/30; **G06F-017/60**

ABSTRACT

PROBLEM TO BE SOLVED: To easily obtain the latest and good information on a product, **technique** , etc., obtaining the evaluation of the highest **rank** as to respective items in respective fields from a **Web site** .

SOLUTION: This system has a **database** server 14 which is connected to a network 11 and provided with a storing means for storing data on the product, **technique** , etc., selected as those obtaining the evaluation of the highest **rank** as for the respective items in the respective fields and data on an information source which displays information on the product, **technique** , etc., and can be accessed via the network 11.

12/5/16   (Item 3 from file: 347)
DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

07156297   **Image available**
INFORMATION PROVIDING METHOD AND BIDDING METHOD

PUB. NO.:   2002-024680 [JP 2002024680 A]
PUBLISHED: January 25, 2002 (20020125)
INVENTOR(s):   KEIDAI YUKUHITO
APPLICANT(s):   FORLIFE EXPRESS KK
APPL. NO.:   2000-203725 [JP 2000203725]
FILED: July 05, 2000 (20000705)
INTL CLASS: **G06F-017/60** ; G06F-013/00

ABSTRACT

PROBLEM TO BE SOLVED: To provide a member to be purchaser with evaluation information and detailed information on an advertisement requester and evaluation information and detailed information on a commodity provided by the advertisement client.

SOLUTION: In this constitution to connect a member side terminal 1, an information control server 2 for one-dimensionally controlling providing of various kinds of information, an advertisement client side terminal 3 opening a shop on an **WEB page** controlled by the information control server 2 and performing commodity sales and advertisement via a communication network N, the information control server 2 executes a **procedure** for transmitting information inputted in a member having an arbitrary **attribute** for information inputted by an information request side member having the same **attribute** , a **procedure** for acquiring response information including word-of- mouth communication evaluation information from a member acquiring transmitted information, a **procedure** for totalizing the acquired word-of-mouth communication evaluation information, **ranking** the information and making the information into data base and a **procedure** for transmitting the response information including **ranking** information made into data base to the information request side member.

12/5/17　　**(Item 4 from file: 347)**
DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

07084930　　**Image available**
TALENT RECOGNITION METHOD AND TALENT RECOGNITION SYSTEM DEVICE

PUB. NO.:　　　2001-312578　[JP 2001312578　A]
PUBLISHED:　　November 09, 2001 (20011109)
INVENTOR(s):　SAKAKIBARA NAOKI
APPLICANT(s):　SAKAKIBARA NAOKI
APPL. NO.:　　2000-128823　[JP 2000128823]
FILED:　　　　April 28, 2000 (20000428)
INTL CLASS:　　**G06F-017/60** ; G09B-019/00

ABSTRACT

PROBLEM TO BE SOLVED: To provide a talent recognition method and a talent
recognition system device which turn up, **rate** and recognize talents of
users found throughout the nation using a telecommunication network, and
which can provide a meeting **site** for the users and enterprises, etc.,
seeking human resources having specific talents.

SOLUTION: This method adopts a **characteristic** constituting **technique**
provided with a registration processing (ST1) which generates and registers
personal data based on the result of responses from a user 4 for plurality
of inquiries; a talent recognition processing (ST2, 3) which judges whether
the personal data satisfies a standard; an approval processing (ST5, 5)
which, when the personal data is judged to satisfy the standard, a talent
recognition council 6 composed of plurality of celebrities makes a judgment
on whether the personal data should be released; and a formal registration
processing (ST6) which, when the approval processing makes a judgment that
the personal data should be released, the personal data is compiled into a
database and is made accessible through the Internet 3.

COPYRIGHT: (C)2001,JPO

12/5/18　　　**(Item 5 from file: 347)**
DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

07079118　　**Image available**
METHOD AND SYSTEM FOR SUPPORTING MERIT **RATING**

PUB. NO.:　　　2001-306764　[JP 2001306764　A]
PUBLISHED:　　November 02, 2001 (20011102)
INVENTOR(s):　FUKUI KOICHI
　　　　　　　　TANIMOTO MITSUNORI
　　　　　　　　TOKUNAGA TAKANARI
APPLICANT(s):　ASAHI BANK RESEARCH INSTITUTE CO LTD
　　　　　　　　PRO HOUSE KK
APPL. NO.:　　2000-117455　[JP 2000117455]
FILED:　　　　April 19, 2000 (20000419)
INTL CLASS:　　**G06F-017/60**

ABSTRACT

PROBLEM TO BE SOLVED: To provide a system for constructing a means for performing merit **rating** on an information communication network, efficiently linking a manager, an evaluator and a person to be evaluated related to merit **rating** and performing consulting for constructing the way of thinking or framework original for a utilizing enterprise corresponding to suggestions or introductions through the relevant system.

SOLUTION: While utilizing a **home page** or the like on the Internet, an enterprise information register means 11 and an employee information register means 13 register various kinds of information concerning the performance of merit **rating** from a manager terminal 3 of the utilizing enterprise on an enterprise information **database** 20, an evaluation item **database** 21 and an employee information **database** 22. After the authentication **procedure** of a utilizing enterprise authenticating means 18 by means of an enterprise ID or the like, an evaluator terminal 4 registers an evaluation point concerning the previously registered person to be evaluated on an analyzed result **database** 23 by means of an evaluation point register means 14. The manager simulates evaluation data registered by an analytic condition register means 16 and an analyzed result read means 17 and obtains merit **rating** data required for assessing a bonus, rise in **rank** or promotion to a higher status.

WITH AN EVALUATION MATRIX

\

WITH AN EVALUATION MATRIX

14/5/1   (Item 1 from file: 2)
DIALOG(R)File   2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

6650206   INSPEC Abstract Number: C2000-08-7250R-045
 Title: Recognizing structure in Web pages using similarity queries
   Author(s): Cohen, W.W.
   Author Affiliation: Shannon Lab., AT&T Bell Labs., Florham Park, NJ, USA
   Conference Title: Proceedings Sixteenth National Conference on Artificial
Intelligence (AAI-99). Eleventh Innovative Applications of Artificial
Intelligence Conference (IAAI-99)   p.59-66
   Publisher: AAAI Press, Menlo Park, CA, USA
   Publication Date: 1999 Country of Publication: USA   xxvi+998 pp.
   ISBN: 0 262 51106 1   Material Identity Number: XX-1999-01742
   Conference Title: Proceedings Sixteenth National Conference on Artificial
Intelligence (AAAI-99). Eleventh Innovative Applications of Artificial
Intelligence Conference (IAAI-99)
   Conference Sponsor: American Assoc. Artificial Intelligence; ACM/SIGART;
Defense Advance Res. Projects Agency; et al
   Conference Date: 18-22 July 1999   Conference Location: Orlando, FL, USA
   Language: English   Document Type: Conference Paper (PA)
   Treatment: Practical (P)
   Abstract: We present general-purpose methods for recognizing certain
types of structure in HTML documents. The methods are implemented using
WHIRL, a "soft" logic that incorporates a notion of textual similarity
developed in the information retrieval community. In an experimental
evaluation on 82 **Web pages** , the structure **ranked** first by our **method**
 is "meaningful", i.e., a structure that was used in a hand-coded
"wrapper", or extraction program, for the page, nearly 70% of the time.
This improves on a value of 50% obtained by an earlier method. With
appropriate background information, the structure-recognition methods
described can also be used to learn a wrapper from examples, or for
maintaining a wrapper as a Web page changes format. In these settings, the
top-ranked structure is meaningful nearly 85% of the time. (20 Refs)
   Subfile: C
   Descriptors: data structures; formal logic; information resources;
Internet; learning systems; pattern recognition; query processing
   Identifiers: Web pages; similarity queries; HTML documents; structure
recognition; information retrieval; WHIRL; formal logic; wrapper; learning
systems
   Class Codes: C7250R (Information retrieval techniques); C7210N (
Information networks); C6120 (File organisation); C4210 (Formal logic);
C6170K (Knowledge engineering techniques)
   Copyright 2000, IEE

14/5/2   (Item 2 from file: 2)
DIALOG(R)File   2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

6242651   INSPEC Abstract Number: C1999-06-7250N-016
   **Title: Reasoning about textual similarity in a Web-based information
access system**
   Author(s): Cohen, W.W.
   Author Affiliation: AT&T Labs.-Res., Florham Park, NJ, USA
   Journal: Autonomous Agents and Multi-Agent Systems   vol.2, no.1   p.
65-86
   Publisher: Kluwer Academic Publishers,
   Publication Date: 1999 Country of Publication: Netherlands

Abstract: The degree to which information sources are pre-processed by
Web based information systems varies greatly. In search engines like
Altavista, little pre-processing is done, while in "knowledge integration"
systems, complex site-specific "wrappers" are used to integrate different
information sources into a common database representation. We describe an
intermediate point between these two models. In our system, information
sources are converted into a highly structured collection of small
fragments of text. Database-like queries to this structured collection of
text fragments are approximated using a novel logic called WHIRL, which
combines inference in the style of deductive databases with **ranked**
 retrieval **methods** from information retrieval (IR). WHIRL allows queries
that integrate information from multiple **Web sites** , without requiring
the extraction and normalization of object identifiers that can be used as
keys; instead, operations that in conventional databases require equality
tests on keys are approximated using IR similarity metrics for text. This
leads to a reduction in the amount of human engineering required to field a
knowledge integration system. Experimental evidence is given showing that
many information sources can be easily modeled with WHIRL, and that
inferences in the logic are both accurate and efficient. (26 Refs)
  Subfile: C
  Descriptors: deductive databases; formal logic; inference mechanisms;
information retrieval; Internet; search engines; text analysis
  Identifiers: textual similarity; Web based information access system;
information source pre-processing; search engines; Altavista; knowledge
integration systems; complex site-specific wrappers; information sources;
common database representation; intermediate point; highly structured
collection; small text fragments; database-like queries; structured
collection; novel logic; WHIRL; inference; deductive databases; ranked
retrieval methods; information retrieval; multiple Web sites; object
identifiers; conventional databases; equality tests; IR similarity metrics;
human engineering
  Class Codes: C7250N (Search engines); C7210N (Information networks);
C6130D (Document processing techniques); C6170K (Knowledge engineering
techniques); C6160K (Deductive databases); C7250R (Information retrieval
techniques); C4210 (Formal logic)
  Copyright 1999, IEE

14/5/3   (Item 3 from file: 2)
DIALOG(R)File   2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

6095691   INSPEC Abstract Number: C9901-7250N-005
  **Title: A Web-based information system that reasons with structured
collections of text**
  Author(s): Cohen, W.W.
  Author Affiliation: AT&T Labs.-Res., Florham Park, NJ, USA
  Conference Title: Proceedings of the Second International Conference on
Autonomous Agents   p.400-7
  Editor(s): Sycara, K.P.; Wooldridge, M.
  Publisher: ACM, New York, NY, USA
  Publication Date: 1998 Country of Publication: USA   xi+478 pp.
  ISBN: 0 89791 983 1   Material Identity Number: XX98-01367

U.S. Copyright Clearance Center Code: 0 89791 983 1/98/5...$5.00
    Conference   Title: Proceedings of 2nd International Conference on
Autonomous Agents
    Conference Sponsor: ACM
    Conference Date: 9-13 May 1998   Conference Location: Minneapolis, MN,
USA
    Language: English   Document Type: Conference Paper (PA)
    Treatment: Practical (P)

    Abstract: The degree to which information sources are pre-processed by
Web-based information systems varies greatly. In search engines like
Altavista, little pre-processing is done, while in "knowledge integration"
systems, complex site-specific "wrappers" are used integrate different
information sources into a common database representation. In this paper,
we describe an intermediate between these two models. In our system,
information sources are converted into a highly structured collection of
small   fragments   of   text. Database-like queries to this structured
collection of text fragments are approximated using a novel logic called
WHIRL (Word-based Heterogeneous Information Retrieval Logic), which
combines inference in the style of deductive databases with **ranked**
 retrieval **methods** from information retrieval (IR). WHIRL allows queries
that integrate information from multiple **Web sites** without requiring
the extraction and normalization of object identifiers that can be used as
keys; instead, operations that in conventional databases require equality
tests on keys are approximated using IR similarity metrics for text. This
leads to a reduction in the amount of human engineering required to field a
knowledge integration system. Experimental evidence is given showing that
many information sources can be easily modeled with WHIRL, and that
inferences in the logic are both accurate and efficient. (21 Refs)
    Subfile: C
    Descriptors: deductive databases; ergonomics; formal logic; full-text
databases; inference mechanisms; information resources; information
retrieval; search engines; software agents; string matching
    Identifiers: World Wide Web-based information system; reasoning;
structured text collections; information source pre-processing; search
engines; knowledge integration systems; site-specific wrappers; common
database representation; text fragments; database-like queries; WHIRL;
Word-based Heterogeneous Information Retrieval Logic; inference; deductive
databases; ranked retrieval methods; object identifiers; key equality tests
; approximated operations; similarity metrics; human engineering
    Class Codes: C7250N (Search engines); C7210N (Information networks);
C6160K (Deductive databases); C7250L (Non-bibliographic retrieval systems);
C6170K (Knowledge engineering techniques); C4210 (Formal logic); C7250R (
Information retrieval techniques); C0240 (Ergonomic aspects of computing)

    14/5/4   (Item 4 from file: 2)

5650906   INSPEC Abstract Number: C9709-7250R-007
    Title: Relevancy ranking of Web pages using shallow parsing
    Author(s): Feinstein, Y.Z.; Goldman, C.V.; Mor, Y.; Rosenschein, J.S.
    Author Affiliation: Dept. of Comput. Sci., Hebrew Univ., Jerusalem,
Israel
    Conference   Title: PADD97   Proceedings   of   the   First   International
Conference on the Practical Application of Knowledge Discovery and Data
Mining   p.125-35
    Publisher: Practical Application Co, Blackpool, UK

Publication Date: 1997 Country of Publication: UK   301 pp.
ISBN: 0 9525554 7 6   Material Identity Number: XX97-00716
Conference Title: Proceedings of the First International Conference on
The Practical Application of Knowledge Discovery and Data Mining PADD 97
Conference Sponsor: CompulogNet; ISL; Lionheart Publishing; Logic
Progamming Associates; et al
Conference Date: 23-25 April 1997   Conference Location: London, UK
Availability: PADD, P.O.Box 137, Blackpool, Lancs. FY2 9UN, UK
Language: English   Document Type: Conference Paper (PA)
Treatment: Practical (P)

Abstract: We present a Phrase-Structure-Grammar **method** to **rank** the
relevancy of **Web pages** . This **method** might be used as an aid to an
information agent, assisting it in understanding the information acquired.
The method could guide the agent when evaluating text by taking into
account the syntactic roles of the words in this text. In cases where the
agent presents its user with results from a search, our system could also
assist this agent by ranking the results according to their relevance to
the user's query. An existing search engine can also take advantage of this
kind of method to improve the precision of its search results, without
undermining recall. In both cases, we focus on ranking the pages retrieved
for the user based on the relevance of these pages to the user's query.
Therefore, we expect that our method will decrease the need to browse
through irrelevant pages. The system is based on a syntactic analysis that
requires the construction of parse trees for part of the text.
Nevertheless, we combine shallow parsing with heuristics, thus making the
system's computational requirements practical for real-world applications.
(12 Refs)
Subfile: C
Descriptors: grammars; heuristic programming; Internet; online front-ends
; query processing; relevance feedback; software agents; trees
(mathematics)
Identifiers: Web page relevancy ranking; shallow parsing; World Wide Web;
Internet; Phrase-Structure-Grammar method; information agent; information
retrieval; syntactic roles; text; searching; search engine; browsing;
syntactic analysis; parse trees; heuristics; computational requirements
Class Codes: C7250R (Information retrieval techniques); C7210 (
Information services and centres); C4210L (Formal languages and
computational linguistics); C6170 (Expert systems); C7250N (Front end
systems for online searching)
Copyright 1997, IEE

14/5/5   (Item 5 from file: 2)
DIALOG(R)File   2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

04320077   INSPEC Abstract Number: C9302-3360L-047
**Title: Maximizing the determinant of the information matrix with the
effective independence method**
Author(s): Poston, W.L.; Tolson, R.H.; Kammer, D.C.
Author Affiliation: George Washington Univ., Hampton, VA, USA
Journal: Journal of Guidance, Control, and Dynamics   vol.15, no.6   p.
1513-14
Publication Date: Nov.-Dec. 1992 Country of Publication: USA
CODEN: JGCDDT ISSN: 0731-5090
Language: English   Document Type: Journal Paper (JP)
Treatment: Theoretical (T)
Abstract: A method has been presented by Kammer (1991) that addresses the
problem of optimally placing sensors on a large space structure for the

purpose of on-orbit modal testing. The **method** **ranks** potential sensor
sites according to their contribution to the independence of the target
modes and iteratively deletes the sites that have the lowest ranking. It is
implied that deleting sensor sites in this way tends to maximize the
determinant of the Fisher information matrix (FIM). The present paper
provides a proof that deleting the potential sensor location with the
smallest effective independence distribution (E/sub D/) value will produce
the smallest relative change in the determinant of the information matrix,
and so this method does provide a local maximization of the determinant of
the FIM. (5 Refs)
   Subfile: C
   Descriptors: aerospace testing; matrix algebra; optimisation
   Identifiers: optimal sensor placement; sensor ranking; optimisation;
aerospace testing; large space structure; on-orbit modal testing; Fisher
information matrix; effective independence distribution
   Class Codes: C3360L (Aerospace systems); C1110 (Algebra); C1180 (
Optimisation techniques)

14/TI/1   (Item 1 from file: 2)
DIALOG(R)File   2:(c) 2005 Institution of Electrical Engineers. All rts. reserv.

  Title: Recognizing structure in Web pages using similarity queries

14/TI/2   (Item 2 from file: 2)
DIALOG(R)File   2:(c) 2005 Institution of Electrical Engineers. All rts. reserv.

  Title: Reasoning about textual similarity in a Web-based information access system

14/TI/3   (Item 3 from file: 2)
DIALOG(R)File   2:(c) 2005 Institution of Electrical Engineers. All rts. reserv.·

  Title: A Web-based information system that reasons with structured collections of text0

14/TI/4   (Item 4 from file: 2)
DIALOG(R)File   2:(c) 2005 Institution of Electrical Engineers. All rts. reserv.

  Title: Relevancy ranking of Web pages using shallow parsing

14/TI/5   (Item 5 from file: 2)
DIALOG(R)File   2:(c) 2005 Institution of Electrical Engineers. All rts. reserv.

  Title: Maximizing the determinant of the information matrix with the effective independence method

14/TI/6   (Item 6 from file: 2)
DIALOG(R)File   2:(c) 2005 Institution of Electrical Engineers. All rts. reserv.

  Title: Site selection for nuclear plants using fuzzy decision analysis

14/TI/7   (Item 7 from file: 2)
DIALOG(R)File   2:(c) 2005 Institution of Electrical Engineers. All rts. reserv.

  Title: Ranking scheme and control token scheme

14/TI/8   (Item 1 from file: 35)
DIALOG(R)File 35:(c) 2005 ProQuest Info&Learning. All rts. reserv.

Solid state nuclear magnetic resonance of membranes

14/TI/9   (Item 2 from file: 35)

Site facilitator roles in videoconferencing: Implications for training

14/TI/10   (Item 3 from file: 35)

DETERMINATION OF HEAD LETTUCE CROP COEFFICIENT AND WATER USE IN CENTRAL ARIZONA (LATUCA SATIVA)

14/TI/11   (Item 4 from file: 35)

A HAZARD RANKING SYSTEM SUITABLE FOR DEVELOPING COUNTRIES   (RISK ASSESSMENT)

14/TI/12   (Item 5 from file: 35)

A METHODOLOGICAL COMPARISON OF NEEDS ASSESSMENT METHODOLOGIES: OBJECTIVE SOCIAL INDICATORS VERSUS THE COMMUNITY SURVEY

14/TI/13   (Item 1 from file: 583)

Chemicals top Infoworld list
   WORLD: CHEMICAL SITES RANK HIGH IN INFOWORLD LISTING

14/TI/14   (Item 1 from file: 6)

Expeditious Methods for Site Characterization and Risk Assessment at Department of Defense Hazardous Waste Sites in the Republic of Korea
   (Master's Thesis)

14/TI/15   (Item 2 from file: 6)

Use of a sensitivity study to identify risk assessment modeling data gaps at the Idaho National Engineering Laboratory's subsurface disposal area

14/TI/16   (Item 3 from file: 6)

Ranking Hazardous-Waste Sites for Remedial Action
   (Final rept)

14/TI/17   (Item 4 from file: 6)

   Risk   Reduction as a Criterion for Measuring Progress of the Installation Restoration Program
   (Master's thesis)

14/TI/18   (Item 5 from file: 6)

   Studies   of coupled chemical and catalytic coal conversion methods. Fifth quarterly report, October--December 1988
   (Progress rept)

14/TI/19   (Item 6 from file: 6)

   Hazardous Waste Site Analysis (Small Site Technology)
   (Final rept. 28 Mar 89-30 Jun 90)

14/TI/20   (Item 7 from file: 6)

   User's Manual for the Defense Priority Model

14/TI/21   (Item 8 from file: 6)

   Installation   Restoration Program. Phase II. Confirmation/Quantification. Stage 1. Homestead Air Force Base, Florida
   (Final rept. Aug 84-Mar 86)

14/TI/22   (Item 9 from file: 6)

   Preliminary Analysis of SOAR Cable Landing Sites at San Clemente Island

14/TI/23   (Item 10 from file: 6)

   Installation   Restoration   Program.   Phase   1.   Records Search, Air Force Plant 44, Tucson, Arizona
   (Final rept)

14/TI/24   (Item 11 from file: 6)

   **Installation Restoration Program Records Search for Air Force Plant 4,
Texas**
   (Final rept. on Phase 1)

14/TI/25   (Item 12 from file: 6)

   **Evaluation and Ranking of Geothermal Resources for Electrical Generation
or Electrical Offset in Idaho, Montana, Oregon and Washington. Volume 1**

14/TI/26   (Item 13 from file: 6)

   **Installation Restoration Program. Phase I. Records Search, Reese, AFB,
Texas**
   (Draft rept)

14/TI/27   (Item 14 from file: 6)

   Hydropower Utilization in New York State
   (Final rept)

14/TI/28   (Item 15 from file: 6)

   **Social Benefits and Assessment of Local Urban Open Space: Recommended
Program of Environ**mental Design Research Priorities (Environmental Design
Research, Problems and Needs)
   (Study rept)

14/TI/29   (Item 16 from file: 6)

   Aquatic Natural Areas in Idaho
   (Research technical completion rept. Dec 74-Dec 76)

14/TI/30   (Item 1 from file: 7)

Title: Corpus-based statistical screening for phrase identification

14/TI/31   (Item 2 from file: 7)

Title: Assessing departures from log-normality in the rank-size rule

14/TI/32   (Item 3 from file: 7)

Title: Skills critical to long-term profitability of engineering firms

14/TI/33   (Item 4 from file: 7)

**Title: LANDFILL SITING USING GEOGRAPHIC INFORMATION-SYSTEMS - A
    DEMONSTRATION**

14/TI/34   (Item 1 from file: 8)

   **Title: Hydro-spatial hierarchical method for siting water harvesting
reservoirs in dry areas**

14/TI/35   (Item 2 from file: 8)

   **Title: Assessment of nine accelerated atmospheric corrosion sites on the
cosmetic corrosion performance of the AISI materials - interim report.**

14/TI/36   (Item 3 from file: 8)

   Title: Small hydro site ranking, cost methodology and program.

14/TI/37   (Item 4 from file: 8)

   Title: Trial of a new Bypass Appraisal Method for lorry nuisance.

14/TI/38   (Item 5 from file: 8)

   Title: USE OF MODEL TESTING IN THE DEVELOPMENT OF SPECIAL SILENCERS.

14/TI/39   (Item 6 from file: 8)

   **Title: RANKING DIRECT USE GEOTHERMAL RESOURCES IN THE NORTHWEST BASED ON
DEVELOPABILITY AND COST.**

14/TI/40   (Item 7 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

   Title: AUSTRALIAN TEST FOR DECAY IN PAINTED TIMBERS EXPOSED TO THE
WEATHER.

14/TI/41   (Item 8 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

  Title: CASE HISTORY OF NUCLEAR POWER PLANT SITE SELECTION.

14/TI/42   (Item 9 from file: 8)
DIALOG(R)File   8:(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

   Title: ALTERNATIVE EVALUATION OF POWER PLANT SITES

14/TI/43   (Item 1 from file: 34)0
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: Evaluating the suitability of habitat for the great crested newt
     (Triturus cristatus)

14/TI/44   (Item 2 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: Regulation of competence development in Haemophilus influenzae -
     Proposed competence regulatory elements are CRP-binding sites

14/TI/45   (Item 3 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: A quantitative ranking of Canada's research output of original human
     studies for the decade 1989 to 1998

14/TI/46   (Item 4 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: Radiation induced sarcoma of the head and neck

14/TI/47   (Item 5 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci. Info. All rts. reserv.

Title: Identification and energetic ranking of possible docking sites for
     pterin on dihydrofolate reductase

14/TI/48   (Item 6 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: Comparison of methods for measuring rabbit incidence on grasslands

 14/TI/49   (Item 7 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: Are Collembola useful as indicators of the conservation value of
    native grasslands?

 14/TI/50   (Item 8 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: MULTICENTER EVALUATION OF 4 METHODS FOR CLOSTRIDIUM-DIFFICILE
    DETECTION - IMMUNOCARD CLOSTRIDIUM-DIFFICILE, CYTOTOXIN ASSAY, CULTURE,
    AND LATEX AGGLUTINATION

 14/TI/51   (Item 9 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: ADRENALECTOMY FOR METASTATIC DISEASE TO THE ADRENAL-GLANDS

 14/TI/52   (Item 10 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: THREATENED STATUS, RARITY, AND DIVERSITY AS ALTERNATIVE SELECTION
    MEASURES FOR PROTECTED AREAS - A TEST USING - A TEST USING AFROTROPICAL
    ANTELOPES

 14/TI/53   (Item 11 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: APPLICATION OF MULTICRITERIA CHOICE-METHODS IN ASSESSING
    EUTROPHICATION

 14/TI/54   (Item 12 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: A SHIFTED MULTIPLICATIVE MODEL CLUSTER-ANALYSIS FOR GROUPING
    ENVIRONMENTS WITHOUT GENOTYPIC RANK CHANGE

 14/TI/55   (Item 13 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: EFFECT OF BODY LOCALE AND ADDITION OF EPINEPHRINE ON THE DURATION OF
    ACTION OF A LOCAL-ANESTHETIC AGENT

 14/TI/56   (Item 14 from file: 34)
DIALOG(R)File 34:(c) 2005 Inst for Sci Info. All rts. reserv.

Title: A SITE SELECTION CASE-STUDY USING TERRAIN ANALYSIS IN CONJUNCTION

10/3,K/1   (Item 1 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

02537030 268883871
Effective techniques for automatic extraction of Web publications
Fong, A C M; Hui, S C; Vu, H L
Online Information Review v26n1 PP: 4-18 2002
ISSN: 1468-4527 JRNL CODE: ONCD
WORD COUNT: 4895

...ABSTRACT: information monitoring systems are ineffective in finding and
tracking scholarly publications. This article analyzes the **characteristics**
of publication index pages and describes effective automatic extraction
techniques that the authors have developed. The authors' **techniques**
combine lexical and syntactic analyses with heuristics. The proposed
techniques have been implemented and tested for more than 14,000 **Web**
pages and achieved consistently high success **rates** of around 90%.
...TEXT: with heuristics. The proposed techniques have been implemented and
tested for more than 14,000 **Web** **pages** and achieved consistently high
success **rates** of around 90 percent.

Introduction

The World Wide Web (WWW) is fast becoming the preferred...
...a search query. Many Web sites that are important to researchers are
omitted in the **process** . Also, most commercial search engines do not index
document files that are in PDF or PS format. It is therefore necessary to
develop **techniques** for automatic extraction of, first of all, research
index Web pages and, second, research papers...

...such as biographies of individuals, on the same Web page. In this paper,
we describe **techniques** and algorithms that have been implemented and
tested to provide effective and automatic extraction of...

...The remainder of the paper is as follows. First, we give an overview of
related **techniques** and systems. Next, we present a thorough analysis of
publication citation Web pages. Effective extraction **techniques** and
algorithms are then presented. We then describe the implementation of a
system that provides...successfully retrieved more than 90 percent of
useful publication information. Unlike approaches that rely on **neural**
networks , our algorithms do not require training/retraining to achieve
useful results.

Appendix

Electronic access
The...

...WIRE - a WWW-based information retrieval and extraction system", 9th
International Workshop on Database and **Expert** **Systems** Applications
(DEXA'98), 26-28 August, Vienna.

AllResearch (2001), "WebClipping Service home page", http:/www...

10/3,K/2   (Item 2 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

02378225   126460601
Web-user satisfaction: An exploratory study
Otto, James R; Najdawi, Mohammad K; Caron, Karen M
Journal of End User Computing v12n4 PP: 3-10 Oct-Dec 2000
ISSN: 1063-2239 JRNL CODE: EUC
WORD COUNT: 5416

...TEXT: Kerlinger's (1986) approach. He cites correlations between total
scores and item scores as a **method** of construct validation. This assumes
that the total score for the survey instrument is valid...

...was created in our survey using three items. The items included were:
"How would you **rate** the overall quality of the **Web** **site** ?", "How well
does the Web site meet your expectations?" and "How satisfied are you with
...

...site?" The extent to which each item and factor is correlated with this
three-item **criterion** scale is a measure of the **criterion** -related
validity. The results of these correlations are presented in **Table** 4.
This **table** shows that several items (C2, F2, E2 and R2) do not have a
meaningful correlation with the overall satisfaction **criterion** . Note,
however, that the correlations between the satisfaction **criterion** and Web
graphics (G 1 = .70 and G2 = .70) are highly significant (p<.000). The
correlation of the entire 12-item construct with the three-item **criterion**
is 0.78, which is significant at p<.000. Please note that the numbers in
parenthesis in **Table** 4 represent measures of significance.

Next, convergent and discriminant validity of this research was assessed...
Systems. His research efforts focus on production scheduling, integration
of IS technologies, and Application of **Expert** **Systems** and AI in
decision making. He was an associate editor for CACM and is an...

...commerce and other MIS courses. His research interests are in the areas
of data mining, **artificial** **intelligence** , and electronic commerce. Dr.
Otto's work has appeared in Annals of OR and International...

 10/3,K/3   (Item 3 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

02325917   86925450
The evolution of Web searching
Green, David
Online Information Review v24n2 PP: 124-137 2000
ISSN: 1468-4527 JRNL CODE: ONCD
WORD COUNT: 8487

...TEXT: Hit (www.directhit.com) represented a radical new departure from
these approaches, and dubbed its **methodology** "the third way". The system
was claimed to be user-controlled as the ranking of...

...Prior to licensing Direct Hit, HotBot returned a list of results based
on the standard **methodology** of matching search terms with content on the
Web sites in its index. Now, Direct...

...identify those Web sites which are popular, according to the number of
visits that each **Web   site** has received, and then re- **rank** the search
results accordingly, with the most popular Web sites that match your search
term...

...and Yahoo!It has now been licensed by AltaVista for its own search site.
However, **artificial   intelligence** (AI) experts have criticised the
company's natural language claims. It was named after the...

 10/3,K/4   (Item 1 from file: 610)
DIALOG(R)File 610:Business Wire
(c) 2005 ·Business Wire. All rts. reserv.

00979676 20031022295B7047 (USE FORMAT 7 FOR FULLTEXT)
**Cerberian's New Filtering and Rating Technologies Provide Most Relevant and
Accurate Web Content Filtering-Proprietary Dynamic Real-Time Rating and
Dynamic Background Rating technologies raise industry standards of Web
filtering accuracy**
Business Wire
Wednesday, October 22, 2003 09:30 EDT
JOURNAL CODE: BW LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 814

...database grows organically
from the surfing habits of its users through our real-time and **artificial
 intelligence** technologies. The result, increased speed and accuracy of
the
content rating **process** . Unlike other filtering services which are one
layer
deep, Cerberian uses a three level **process** consisting of first, a
relevant
database of rated and categorized sites and domains; second, Cerberian's
Dynamic Real-Time Rating (DRTR) technology, and finally, Cerberian's
Dynamic
Background Rating (DBR) **process** .

"Our **process** ensures that our users are the most protected and safe users
on
the Internet benefiting...

...seamlessly to the user.

Dynamic Real-Time Rating / Dynamic Background Rating

The Dynamic Real-Time **Rating** (DRTR) technology retrieves non- **rated   Web
  pages**
from their host servers to be analyzed for content. Cerberian's DRTR
technology looks at....

...context for
each word, links to and from the page, and other sophisticated page
analyzing
 techniques and then responds in one of two ways. **Web   sites** that can
be **rated**
as adult material, pornography or other high-liability categories are
rated,

categorized and immediately blocked...

...on the user's policy.
These sites are then categorized in Cerberian's master ratings **database**
insuring that every other Cerberian user worldwide has the same, up-to-date
information. Sites that fall into other categories are moved to Cerberian's
DBR **process** for additional review.

Cerberian's Dynamic Background Rating (DBR) **process** uses a number of
proprietary ratings modules that individually rate and categorize a page
based...

 10/3,K/5   (Item 2 from file: 610)
DIALOG(R)File 610:Business Wire
(c) 2005 Business Wire. All rts. reserv.

00660213 20020206037B8535 (USE FORMAT 7 FOR FULLTEXT)
**Technology Catches Up to Internet Filtering Software Industry; Cerberian's
Internet Manager Delivers Fast and Reliable Web-Based Filtering Software to
Businesses**
Business Wire
Wednesday, February 6, 2002 13:24 EST
JOURNAL CODE: BW LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 784

...expensive software and hardware. Cerberian's Internet Manager is
designed to quicken the Internet filtering **process** while making it more
cost
effective for companies to implement monitoring and filtering policies."

Cerberian...

...Unlike many of today's filtering software applications, Cerberian does
not
rely exclusively on a **database** of pre-viewed and categorized URLs.
Cerberian's
DRTR service, based on **Neural   Net (artificial   intelligence)**
technology, is an
embedded and online service that dynamically reads, **rates** and categorizes
Web
sites on the fly.

If the URL entered by the user is not in the database...

...appeared in
the early 1990s," Smith said.

"Along with nearly eliminating the slow manual reviewing **process** ,
Cerberian
has lowered the costs, reduced maintenance and developed a solution for
accurately reading and...

INDUSTRY NAMES: **ARTIFICIAL   INTELLIGENCE ;**

 10/3,K/6   (Item 3 from file: 610)

15/3,K/3   (Item 3 from file: 9)
DIALOG(R)File   9:Business & Industry(R)
(c) 2005 The Gale Group. All rts. reserv.

01343068 Supplier Number: 23996871   (USE FORMAT 7 OR 9 FOR FULLTEXT)
Software Agent VARs Being Scoped By IBM
(IBM is targeting VARs with new software agent technology; commercial
  version of agent software, called WBI Personal Agent, is being licensed
  by IBM)
Computer Reseller News, p 69
August 18, 1997
DOCUMENT TYPE: Journal ISSN: 0893-8377 (United States)
LANGUAGE: English RECORD TYPE: Fulltext
WORD COUNT: 229

      (USE FORMAT 7 OR 9 FOR FULLTEXT)

TEXT:
...the site, provide alerts to speeds of links, advise the user to changes
at a **Web   site** , **rank** order of viewed sites by frequency and **how**
recently they have been visited, and learn user patterns and suggest
shortcuts, IBM executives said...

15/3,K/5   (Item 2 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02446340   SUPPLIER NUMBER: 65859280   (USE FORMAT 7 OR 9 FOR FULL TEXT)
Microsoft Serves Up ISA.(Software Review)(Evaluation)
MCFADDEN, MARK
ENT, 5, 15, 36
Sept 20, 2000
DOCUMENT TYPE: Evaluation   ISSN: 1085-2395   LANGUAGE: English
RECORD TYPE: Fulltext; Abstract
WORD COUNT: 1179   LINE COUNT: 00097

...   user requests -- and the far more effective active cache. We set up
ISA Server to **rank** the most commonly visited **Web   sites** , determine
how often those sites update their content, and then automatically obtain
and cache new content when...

15/3,K/6   (Item 3 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02408065   SUPPLIER NUMBER: 62652909   (USE FORMAT 7 OR 9 FOR FULL TEXT)
A Community of Favorites.(A Community of Favorites -- Quiver.com gleans
  your favorite sites to build a best-of-Web directory, with a few added
  bonuses, but beware what privacy you give up.)(Software
  Review)(Evaluation)
Bannan, Karen J.
WinMag.com, NA
April 26, 2000
DOCUMENT TYPE: Evaluation   LANGUAGE: English   RECORD TYPE: Fulltext
  ; Abstract
WORD COUNT: 933   LINE COUNT: 00071

TEXT:
     ...via the Qbar's Friends button Of course, the site's real draw is
its **Web site** aggregation and **ranking** system. You can see **how**
popular a site is by looking at the small people icons that sit next to...

   15/3,K/7   (Item 4 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02406353   SUPPLIER NUMBER: 62695117   (USE FORMAT 7 OR 9 FOR FULL TEXT)
**Content Matters Most In Search-Engine Placement -- WITH MORE ENGINES
   RELYING ON HUMAN EDITORS, IT'S WHAT A SITE CONTAINS THAT
   COUNTS.(Internet/Web/Online Service Information)**
Kahaner, Larry
InformationWeek, 172
June 12, 2000
ISSN: 8750-6874   LANGUAGE: English   RECORD TYPE: Fulltext; Abstract
WORD COUNT: 2202   LINE COUNT: 00169

...   or arcane tactics. It takes Web-design skills, perseverance, hard
work, a thorough knowledge of **how** the various search engines **rank Web
sites** , a smattering of good luck, and, most of all, compelling content.
     In fact, now that...

   15/3,K/8   (Item 5 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02392019   SUPPLIER NUMBER: 61602912   (USE FORMAT 7 OR 9 FOR FULL TEXT)
WEB SEARCHER'S COMPANION.(Buyers Guide)
Negrino, Tom
Macworld, 17, 5, 76
May, 2000
DOCUMENT TYPE: Buyers Guide   ISSN: 0741-8647   LANGUAGE: English
RECORD TYPE: Fulltext
WORD COUNT: 4324   LINE COUNT: 00362

...   or a string of words, and the site searches its index for those
specific words. **Web - site** designers can affect **how** highly their site
is **ranked** by search engines, with careful selection of page titles, body
copy, and invisible HTML tags...

   15/3,K/9   (Item 6 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02342182   SUPPLIER NUMBER: 56471937   (USE FORMAT 7 OR 9 FOR FULL TEXT)
Net Manners Matter: How **Top Sites** Rank **in Social Behavior.(four popular
   Web   sites)(Company Business and Marketing)**
Computerworld, 40(1)
Oct 18, 1999
ISSN: 0010-4841   LANGUAGE: English   RECORD TYPE: Fulltext; Abstract
WORD COUNT: 1639   LINE COUNT: 00136

Net Manners Matter: How **Top Sites** Rank **in Social Behavior.**(four popular
    Web   sites)(Company **Business and Marketing)**

  15/3,K/13   (Item 10 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02078391   SUPPLIER NUMBER: 19550193   (USE FORMAT 7 OR 9 FOR FULL TEXT)
Internet Update 06/30/97:The Latest News On Search Engines.
Newsbytes, pNEW06300024
June 30, 1997
LANGUAGE: English   RECORD TYPE: Fulltext
WORD COUNT: 72   LINE COUNT: 00008

TEXT:
      ...for "A Webmaster's Guide To Search Engines." The new site include
additional material including **how** search engines **rank Web pages** ,
how long you'll have to wait for a listing after submission, how
frequently the search...

  15/3,K/14   (Item 11 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

01961504   SUPPLIER NUMBER: 18483203   (USE FORMAT 7 OR 9 FOR FULL TEXT)
Navigating the Internet: Quarterdeck's WebCompass eases Web information
  management. (WebCompass 1.02 search tool) (Lab Note) (Software
  Review)(Brief Article)(Evaluation)
Rapoza, Jim
PC Week, v13, n28, p43(2)
July 15, 1996
DOCUMENT TYPE: Brief Article Evaluation   ISSN: 0740-1604
LANGUAGE: English   RECORD TYPE: Fulltext
WORD COUNT: 599   LINE COUNT: 00048

...   comments. Afterward, if we needed to return to the topic, we could
easily sort the **Web   pages** by **rankings** .
      We were able to configure **how** often the agent updated page summaries
and topics as well as when to time out...

  15/3,K/18   (Item 4 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

01677716   Supplier Number: 50170944   (USE FORMAT 7 FOR FULLTEXT)
WebPostion Gold 1.0: First Web Site Promotion Software To Shift Balance of
  Power From Search Engines to Marketers.
Business Wire, p07150266
July 15, 1998
Language: English   Record Type: Fulltext
Article Type: Article
Document Type: Newswire; Trade
Word Count: 579

...   found near the top of the results, it might as well be invisible,"

said Winters. " **How** search engines determine where a **Web site ranks**
has long been their most carefully guarded secret. If a Web marketer is
willing to...

 15/3,K/19   (Item 5 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

01627209   Supplier Number: 48379130 (USE FORMAT 7 FOR FULLTEXT)
**A New Internet Agent, ScoreCheck, Determines Competitive Website Rankings
   On Search Engines In A Volatile Online Advertising Market.**
Business Wire, p03261251
March 26, 1998
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 429

   (USE FORMAT 7 FOR FULLTEXT)
TEXT:
...BUSINESS WIRE)--March 26, 1998--ScoreCard Inc. Thursday announced
ScoreCheck, its automated agent that analyzes **how web sites rank**
with 15 major internet search engines.

 15/3,K/20   (Item 6 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

01577629   Supplier Number: 48035890 (USE FORMAT 7 FOR FULLTEXT)
**Submit It! Acquires PositionAgent from NetGambit, Delivering Significant
   Competitive Advantage to Customers**
PR Newswire, p1006NEM042
Oct 6, 1997
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 648

...   our ongoing effort to aggressively meet that need," he said.
   "Now customers can easily check **how** their **Web sites rank** and
implement strategies for improving their rankings -- all from the same
vendor," said Younker. "This...

 15/3,K/21   (Item 7 from file: 621)
DIALOG(R)File 621:Gale Group New Prod.Annou.(R)
(c) 2005 The Gale Group. All rts. reserv.

01555084   Supplier Number: 47869243 (USE FORMAT 7 FOR FULLTEXT)
IBM Software Agent Technology Helps Users Control Web Information.
Business Wire, p07301208
July 30, 1997
Language: English   Record Type: Fulltext
Document Type: Newswire; Trade
Word Count: 577

...   the site, provide alerts to speeds of links, advise the user to
changes at a **web site , rank** order viewed sites by frequency and **how**

recently they have been visited, learn user patterns and suggest shortcuts.
        For enterprise customers, software...

 15/3,K/22   (Item 1 from file: 636)
DIALOG(R)File 636:Gale Group Newsletter DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

.04639573   Supplier Number: 61700620   (USE FORMAT 7 FOR FULLTEXT)
New service tracks web site rankings in search portals.
Telecomworldwire, pNA
April 26, 2000
Language: English   Record Type: Fulltext
Document Type: Newsletter; Trade
Word Count: 110

...   search engine checker that is designed to enable webmasters and
search engine specialists to monitor **how** their **web   site   ranks**
within search engines has been launched.
        Called www.whatwhywhere.com the service tracks search engine...

· **15/3,K/33   (Item 1 from file: 148)**
DIALOG(R)File 148:Gale Group Trade & Industry DB
(c)2005 The Gale Group. All rts. reserv.

11910972   SUPPLIER NUMBER: 59959870   (USE FORMAT 7 OR 9 FOR FULL TEXT)
Finding information by finding search engines.
WATT,· MICHAEL
LI Business News, 47, 6, 33A
Feb 11, 2000
ISSN: 0894-4806   LANGUAGE: English   RECORD TYPE: Fulltext
WORD COUNT: 891   LINE COUNT: 00068

...   you are looking appears within the Web site. To learn more about
search engines -- particularly **how** you register with them and **how** they
rank   web   pages -- and to stay abreast of new developments in search
engine technology, visit www.searchenginewatch.com...

 15/3,K/38   (Item 6 from file: 148)
DIALOG(R)File 148:Gale Group Trade & Industry DB
(c)2005 The Gale Group. All rts. reserv.

07977716   SUPPLIER NUMBER: 17187876   (USE FORMAT 7 OR 9 FOR FULL TEXT)
**Web sites for cybernauts.(online services guide)(includes related articles
  on popular World Wide Web sites and software)**
Flynn, Mary Kathleen
U.S. News & World Report, v119, n2, p48(4)
July 10, 1995
ISSN: 0041-5537   LANGUAGE: English   RECORD TYPE: Fulltext; Abstract
WORD COUNT: 2673   LINE COUNT: 00224

...   Kathleen Flynn can be reached via E-mail at
73552.3326@compuserve.com
        The hot **Web   sites**
        Ranked by "hits"-- **how** often any file in a **Web   site** is accessed
by a user--these sites are currently the most popular on the Web...

15/TI/1   (Item 1 from file: 9)
DIALOG(R)File   9:(c) 2005 The Gale Group. All rts. reserv.

Banks Adopt Net-Based Bill Payment Services

15/TI/2   (Item 2 from file: 9)
DIALOG(R)File   9:(c) 2005 The Gale Group. All rts. reserv.

Banks Held Back on Bill Presentment Front, But Make Progress in Other Areas

15/TI/3   (Item 3 from file: 9)
DIALOG(R)File   9:(c) 2005 The Gale Group. All rts. reserv.

Software Agent VARs Being Scoped By IBM

15/TI/4   (Item 1 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

**Pings&Packets - Searching the industry for technical connections and
   returning analysis in byte-size packages.(Industry Trend or
   Event)(Column)**

15/TI/5   (Item 2 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

Microsoft Serves Up ISA.(Software Review)(Evaluation)

15/TI/6   (Item 3 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

**A Community of Favorites.(A Community of Favorites -- Quiver.com gleans
   your favorite sites to build a best-of-Web directory, with a few added
   bonuses, but beware what privacy you give up.)(Software
   Review)(Evaluation)**

15/TI/7   (Item 4 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

**Content Matters Most In Search-Engine Placement -- WITH MORE ENGINES
   RELYING ON HUMAN EDITORS, IT'S WHAT A SITE CONTAINS THAT
   COUNTS.(Internet/Web/Online Service Information)**

1 **15/TI/8   (Item 5 from file: 275)**
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

WEB SEARCHER'S COMPANION.(Buyers Guide)

15/TI/9   (Item 6 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

Net Manners Matter: How **Top Sites** Rank **in Social Behavior.**(four popular
   Web   sites)(**Company Business and Marketing**)

15/TI/10   (Item 7 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

IT.com Supersites.(Web sites are aiding the IT procurement
   process)(Industry Trend or Event)

15/TI/11   (Item 8 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

Bargain-basement PCs make the perfect tools for training
   employees.(Emachines eTower PC)(Hardware Review)(Evaluation)

15/TI/12   (Item 9 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

There's Money to Be Made in Online Customer Data.(Industry Trend or Event)

15/TI/13   (Item 10 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

Internet Update 06/30/97:The Latest News On Search Engines.

15/TI/14   (Item 11 from file: 275)
DIALOG(R)File 275:(c) 2005 The Gale Group. All rts. reserv.

Navigating the Internet: Quarterdeck's WebCompass eases Web information
   management. (WebCompass 1.02 search tool) (Lab Note) (Software
   Review)(Brief Article)(Evaluation)

15/TI/15   (Item 1 from file: 621)
DIALOG(R)File 621:(c) 2005 The Gale Group. All rts. reserv.

Newstream.com digest: Ranking **the Airline** Websites , How **Not to Buy a**
   Lemon Other Multimedia for Journalists.

15/TI/16   (Item 2 from file: 621)
DIALOG(R)File 621:(c) 2005 The Gale Group. All rts. reserv.

etown.com Ranked No. 1 Overall Online Electronics Store.

15/TI/17   (Item 3 from file: 621)
DIALOG(R)File 621:(c) 2005 The Gale Group. All rts. reserv.

Art Technology Group's Dynamo Deployed At jcrew.com; World-Class Retail Web
  Site Presents Unique Online Customer Experience.

 15/TI/18   (Item 4 from file: 621)
DIALOG(R)File 621:(c) 2005 The Gale Group. All rts. reserv.

WebPostion Gold 1.0: First Web Site Promotion Software To Shift Balance of
  Power From Search Engines to Marketers.

 15/TI/19   (Item 5 from file: 621)
DIALOG(R)File 621:(c) 2005 The Gale Group. All rts. reserv.

A New Internet Agent, ScoreCheck, Determines Competitive Website Rankings
  On Search Engines In A Volatile Online Advertising Market.

 15/TI/20   (Item 6 from file: 621)
DIALOG(R)File 621:(c) 2005 The Gale Group. All rts. reserv.

Submit It! Acquires PositionAgent from NetGambit, Delivering Significant
  Competitive Advantage to Customers

 15/TI/21   (Item 7 from file: 621)
DIALOG(R)File 621:(c) 2005 The Gale Group. All rts. reserv.

IBM Software Agent Technology Helps Users Control Web Information.

 15/TI/22   (Item 1 from file: 636)
DIALOG(R)File 636:(c) 2005 The Gale Group. All rts. reserv.

New service tracks web site rankings in search portals.

 15/TI/23   (Item 2 from file: 636)
DIALOG(R)File 636:(c) 2005 The Gale Group. All rts. reserv.

C.W. Henderson Publisher Among Internet's Top Five Sites for Key Health
  Topics.

 15/TI/24   (Item 3 from file: 636)
DIALOG(R)File 636:(c) 2005 The Gale Group. All rts. reserv.

Web of the Week.

 15/TI/25   (Item 4 from file: 636)
DIALOG(R)File 636:(c) 2005 The Gale Group. All rts. reserv.

IBM: IBM software agent technology helps users control web information

 15/TI/26   (Item 5 from file: 636)

3. SOFTWARE MEASURES USERS' PERCEPTION

15/TI/27   (Item 1 from file: 16)

FAVORITE FREEBIES.

15/TI/28   (Item 2 from file: 16)

A Web Search Trifecta.

15/TI/29   (Item 3 from file: 16)

CALL FOR ISPs:WBI Personal Agent: Software Agent VARs Being Scoped By IBM

15/TI/30   (Item 4 from file: 16)

NEW IBM AGENT TECHNOLOGY MAKES IT EASIER TO FIND, CONTROL WEB INFORMATION

15/TI/31   (Item 5 from file: 16)

Internet visitors' traffic jam makes buyers Web wary

15/TI/32   (Item 6 from file: 16)

**Navigating the Internet; Quarterdeck's WebCompass eases Web information
  management**

15/TI/33   (Item 1 from file: 148)

Finding information by finding search engines.

15/TI/34   (Item 2 from file: 148)

**New Year's resolutions for the Web user; recommendations for everyone on
  how to be a savvier surfer in 1998.**

15/TI/35   (Item 3 from file: 148)

Internet Librarian, LibTech International.

15/TI/36   (Item 4 from file: 148)
DIALOG(R)File 148:(c)2005 The Gale Group. All rts. reserv.

Evaluating Net Evaluators.

15/TI/37   (Item 5 from file: 148)
DIALOG(R)File 148:(c)2005 The Gale Group. All rts. reserv.

Net relations: a fusion of direct marketing and public relations.

15/TI/38   (Item 6 from file: 148)
DIALOG(R)File 148:(c)2005 The Gale Group. All rts. reserv.

**Web sites for cybernauts.(online services guide)(includes related articles
on popular World Wide Web sites and software)**

Dialⸯg DataStarⵧ

Document

Select the documents you wish to save or order by clicking the box next to the document,
or click the link above the document to order directly.

save locally as: |PDF document ▾| search strategy: |do not include the search strategy ▾|

previous documents | next documents | order

Fulltext-Link: USPTO Full Text Retrieval Options

☑ **document 7 of 82** Order Document

Examiners' Electronic Digest Database (EEDD)

Accession number & update
0000000279 20050815.
Title
The Anatomy of a Large-Scale Hypertextual Web Search Engine.
Publication Information
Brin, Sergey; Page, Lawrence. The Anatomy of a Large-Scale Hypertextual Web Search Engine.
Computer Networks and ISDN Systems, vol. 30, no. 1-7, April 1998. p. 107-117.
Source
Computer-Networks-and-ISDN-Systems, vol. 30, no. 1-7, p. 107-117, April 1998.
ISSN 01697552.
Document type
article.
Record type
abstract, full text.
Publication date
199804.
Journal title
Computer-Networks-and-ISDN-Systems.
Author(s)
Brin-Sergey; Page-Lawrence.
Link
http://www-db.stanford.edu/~backrub/google.html.
Document identifier
ISSN 01697552.
US classification
705/26;
705/1;
705/14;
705/10.
Examiner comments
Highly relevant. Collaborative Filterings Research Papers includes abstracts from several papers.
Keywords
social-filtering; collaborating-filtering; implicit-ratings; refer; recommend; users-information; contents-
based-filtering; world-wide-web; search-engines; information-retrieval; pagerank; google.

EPO XP number
ECLA class
Document source
Submitting TC
Copyright
Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/ To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in- depth description of our large-scale web search engine -- the first such detailed public description we know of to date. Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Full text

1. Introduction

(Note: There are two versions of this paper -- a longer full version and a shorter printed version. The full version is available on the web and the conference CD-ROM.)

The web creates new challenges for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. People are likely to surf the web using its link graph, often starting with high quality human maintained indices such as Yahoo! or with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead automated search engines. We have built a large-scale search engine which addresses many of the problems of existing systems. It makes especially heavy use of the additional structure present in hypertext to provide much higher quality search results. We chose our system name, Google, because it is a common spelling of googol, or 10100 and fits well with our goal of building very large-scale search engines.

1.1 Web Search Engines -- Scaling Up: 1994 - 2000

Search engine technology has had to scale dramatically to keep up with the growth of the web. In 1994, one of the first web search engines, the World Wide Web Worm (WWWW) (McBryan 94) had an index of 110,000 web pages and web accessible documents. As of November, 1997, the top search engines claim to index from 2 million (WebCrawler) to 100 million web documents (from Search Engine Watch). It is foreseeable that by the year 2000, a comprehensive index of the Web will contain over a billion documents. At the same time, the number of queries search engines handle has grown incredibly too. In March and April 1994, the World Wide Web Worm received an average of about 1500 queries per day. In November 1997, Altavista claimed it handled roughly 20 million queries per day. With the increasing number of users on the web, and automated systems which query search engines, it is

likely that top search engines will handle hundreds of millions of queries per day by the year 2000. The goal of our system is to address many of the problems, both in quality and scalability, introduced by scaling search engine technology to such extraordinary numbers.

1.2. Google: Scaling with the Web

Creating a search engine which scales even to today's web presents many challenges. Fast crawling technology is needed to gather the web documents and keep them up to date. Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently. Queries must be handled quickly, at a rate of hundreds to thousands per second.

These tasks are becoming increasingly difficult as the Web grows. However, hardware performance and cost have improved dramatically to partially offset the difficulty. There are, however, several notable exceptions to this progress such as disk seek time and operating system robustness. In designing Google, we have considered both the rate of growth of the Web and technological changes. Google is designed to scale well to extremely large data sets. It makes efficient use of storage space to store the index. Its data structures are optimized for fast and efficient access (see section 4.2). Further, we expect that the cost to index and store text or HTML will eventually decline relative to the amount that will be available (see Appendix B). This will result in favorable scaling properties for centralized systems like Google.

1.3 Design Goals

1.3.1 Improved Search Quality

Our main goal is to improve the quality of web search engines. In 1994, some people believed that a complete search index would make it possible to find anything easily. According to Best of the Web 1994 -- Navigators, "The best navigation service should make it easy to find almost anything on the Web (once all the data is entered)." However, the Web of 1997 is quite different. Anyone who has used a search engine recently, can readily testify that the completeness of the index is not the only factor in the quality of search results. "Junk results" often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results). One of the main causes of this problem is that the number of documents in the indices has been increasing by many orders of magnitude, but the user's ability to look at documents has not. People are still only willing to look at the first few tens of results. Because of this, as the collection size grows, we need tools that have very high precision (number of relevant documents returned, say in the top tens of results). Indeed, we want our notion of "relevant" to only include the very best documents since there may be tens of thousands of slightly relevant documents. This very high precision is important even at the expense of recall (the total number of relevant documents the system is able to return). There is quite a bit of recent optimism that the use of more hypertextual information can help improve search and other applications (Marchiori 97) (Spertus 97) (Weiss 96) (Kleinberg 98). In particular, link structure (Page 98) and link text provide a lot of information for making relevance judgments and quality filtering. Google makes use of both link structure and anchor text (see Sections 2.1 and 2.2).

1.3.2 Academic Search Engine Research

Aside from tremendous growth, the Web has also become increasingly commercial over time. In 1993, 1.5% of web servers were on .com domains. This number grew to over 60% in 1997. At the same time, search engines have migrated from the academic domain to the commercial. Up until now most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art and to be advertising oriented (see Appendix A). With Google, we have a strong

goal to push more development and understanding into the academic realm.

Another important design goal was to build systems that reasonable numbers of people can actually use. Usage was important to us because we think some of the most interesting research will involve leveraging the vast amount of usage data that is available from modern web systems. For example, there are many tens of millions of searches performed every day. However, it is very difficult to get this data, mainly because it is considered commercially valuable.

Our final design goal was to build an architecture that can support novel research activities on large-scale web data. To support novel research uses, Google stores all of the actual documents it crawls in compressed form. One of our main goals in designing Google was to set up an environment where other researchers can come in quickly, process large chunks of the web, and produce interesting results that would have been very difficult to produce otherwise. In the short time the system has been up, there have already been several papers using databases generated by Google, and many others are underway. Another goal we have is to set up a Spacelab-like environment where researchers or even students can propose and do interesting experiments on our large-scale web data.

2. System Features

The Google search engine has two important features that help it produce high precision results. First, it makes use of the link structure of the Web to calculate a quality ranking for each web page. This ranking is called PageRank and is described in detail in (Page 98). Second, Google utilizes link to improve search results.

2.1 PageRank: Bringing Order to the Web

The citation (link) graph of the web is an important resource that has largely gone unused in existing web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow rapid calculation of a web page's "PageRank", an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, PageRank is an excellent way to prioritize the results of web keyword searches. For most popular subjects, a simple text matching search that is restricted to web page titles performs admirably when PageRank prioritizes the results (demo available at google.stanford.edu). For the type of full text searches in the main Google system, PageRank also helps a great deal.

2.1.1 Description of PageRank Calculation

Academic citation literature has been applied to the web, largely by counting citations or backlinks to a given page. This gives some approximation of a page's importance or quality. PageRank extends this idea by not counting links from all pages equally, and by normalizing by the number of links on a page.

PageRank is defined as follows:

We assume page A has pages $T_1...T_n$ which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d (PR(T_1)/C(T_1) + ... + PR(T_n)/C(T_n))$$

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. Also, a PageRank for 26 million web pages can be computed in a few hours on a medium size workstation. There are many other details which are beyond the scope of this paper.

2.1.2 Intuitive Justification

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps

clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the d damping factor is the probability at each page the "random surfer" will get bored and request another random page. One important variation is to only add the damping factor d to a single page, or a group of pages. This allows for personalization and can make it nearly impossible to deliberately mislead the system in order to get a higher ranking. We have several other extensions to PageRank, again see (Page 98).

Another intuitive justification is that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Intuitively, pages that are well cited from many places around the web are worth looking at. Also, pages that have perhaps only one citation from something like the Yahoo! homepage are also generally worth looking at. If a page was not high quality, or was a broken link, it is quite likely that Yahoo's homepage would not link to it. PageRank handles both these cases and everything in between by recursively propagating weights through the link structure of the web.

2.2 Anchor Text

The text of links is treated in a special way in our search engine. Most search engines associate the text of a link with the page that the link is on. In addition, we associate it with the page the link points to. This has several advantages. First, anchors often provide more accurate descriptions of web pages than the pages themselves. Second, anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programs, and databases. This makes it possible to return web pages which have not actually been crawled. Note that pages that have not been crawled can cause problems, since they are never checked for validity before being returned to the user. In this case, the search engine can even return a page that never actually existed, but had hyperlinks pointing to it. However, it is possible to sort the results, so that this particular problem rarely happens.

This idea of propagating anchor text to the page it refers to was implemented in the World Wide Web Worm (McBryan 94) especially because it helps search non-text information, and expands the search coverage with fewer downloaded documents. We use anchor propagation mostly because anchor text can help provide better quality results. Using anchor text efficiently is technically difficult because of the large amounts of data which must be processed. In our current crawl of 24 million pages, we had over 259 million anchors which we indexed.

2.3 Other Features

Aside from PageRank and the use of anchor text, Google has several other features. First, it has location information for all hits and so it makes extensive use of proximity in search. Second, Google keeps track of some visual presentation details such as font size of words. Words in a larger or bolder font are weighted higher than other words. Third, full raw HTML of pages is available in a repository.

3 Related Work

Search research on the web has a short and concise history. The World Wide Web Worm (WWWW) (McBryan 94) was one of the first web search engines. It was subsequently followed by several other academic search engines, many of which are now public companies. Compared to the growth of the Web and the importance of search engines there are precious few documents about recent search engines (Pinkerton 94). According to Michael Mauldin (chief scientist, Lycos Inc) (Mauldin), "the various services (including Lycos) closely guard the details of these databases". However, there has been a fair amount of work on specific features of search engines. Especially well represented is work which can get results by post-processing the results of existing commercial search engines, or produce small scale "individualized" search engines. Finally, there has been a lot of research on information retrieval systems, especially on well controlled collections. In the next

two sections, we discuss some areas where this research needs to be extended to work better on the web.

3.1 Information Retrieval

Work in information retrieval systems goes back many years and is well developed (Witten 94). However, most of the research on information retrieval systems is on small well controlled homogeneous collections such as collections of scientific papers or news stories on a related topic. Indeed, the primary benchmark for information retrieval, the Text Retrieval Conference (TREC 96), uses a fairly small, well controlled collection for their benchmarks. The "Very Large Corpus" benchmark is only 20GB compared to the 147GB from our crawl of 24 million web pages. Things that work well on TREC often do not produce good results on the web. For example, the standard vector space model tries to return the document that most closely approximates the query, given that both query and document are vectors defined by their word occurrence. On the web, this strategy often returns very short documents that are the query plus a few words. For example, we have seen a major search engine return a page containing only "Bill Clinton Sucks" and picture from a "Bill Clinton" query. Some argue that on the web, users should specify more accurately what they want and add more words to their query. We disagree vehemently with this position. If a user issues a query like "Bill Clinton" they should get reasonable results since there is a enormous amount of high quality information available on this topic. Given examples like these, we believe that the standard information retrieval work needs to be extended to deal effectively with the web.

3.2 Differences Between the Web and Well.Controlled Collections

The web is a vast collection of completely uncontrolled heterogeneous documents. Documents on the web have extreme variation internal to the documents, and also in the external meta information that might be available. For example, documents differ internally in their language (both human and programming), vocabulary (email addresses, links, zip codes, phone numbers, product numbers), type or format (text, HTML, PDF, images, sounds), and may even be machine generated (log files or output from a database). On the other hand, we define external meta information as information that can be inferred about a document, but is not contained within it. Examples of external meta information include things like reputation of the source, update frequency, quality, popularity or usage, and citations. Not only are the possible sources of external meta information varied, but the things that are being measured vary many orders of magnitude as well. For example, compare the usage information from a major homepage, like Yahoo's which currently receives millions of page views every day with an obscure historical article which might receive one view every ten years. Clearly, these two items must be treated very differently by a search engine.

Another big difference between the web and traditional well controlled collections is that there is virtually no control over what people can put on the web. Couple this flexibility to publish anything with the enormous influence of search engines to route traffic and companies which deliberately manipulating search engines for profit become a serious problem. This problem that has not been addressed in traditional closed information retrieval systems. Also, it is interesting to note that metadata efforts have largely failed with web search engines, because any text on the page which is not directly represented to the user is abused to manipulate search engines. There are even numerous companies which specialize in manipulating search engines for profit.

4 System Anatomy

First, we will provide a high level discussion of the architecture. Then, there is some in-depth descriptions of important data structures. Finally, the major applications: crawling, indexing, and searching will be examined in depth.

Figure 1. High Level Google Architecture (image omitted)

4.1 Google Architecture Overview

In this section, we will give a high level overview of how the whole system works as pictured in Figure 1. Further sections will discuss the applications and data structures not mentioned in this section. Most of Google is implemented in C or C++ for efficiency and can run in either Solaris or Linux.

In Google, the web crawling (downloading of web pages) is done by several distributed crawlers. There is a URLserver that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the storeserver. The storeserver then compresses and stores the web pages into a repository. Every web page has an associated ID number called a docID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions. It reads the repository, uncompresses the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link.

The URLresolver reads the anchors file and converts relative URLs into absolute URLs and in turn into docIDs. It puts the anchor text into the forward index, associated with the docID that the anchor points to. It also generates a database of links which are pairs of docIDs. The links database is used to compute PageRanks for all the documents.

The sorter takes the barrels, which are sorted by docID (this is a simplification, see Section 4.2.5), and resorts them by wordID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of wordIDs and offsets into the inverted index. A program called DumpLexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used by the searcher. The searcher is run by a web server and uses the lexicon built by DumpLexicon together with the inverted index and the PageRanks to answer queries.

4.2 Major Data Structures

Google's data structures are optimized so that a large document collection can be crawled, indexed, and searched with little cost. Although, CPUs and bulk input output rates have improved dramatically over the years, a disk seek still requires about 10 ms to complete. Google is designed to avoid disk seeks whenever possible, and this has had a considerable influence on the design of the data structures.

4.2.1 BigFiles

BigFiles are virtual files spanning multiple file systems and are addressable by 64 bit integers. The allocation among multiple file systems is handled automatically. The BigFiles package also handles allocation and deallocation of file descriptors, since the operating systems do not provide enough for our needs. BigFiles also support rudimentary compression options.

4.2.2 Repository

Figure 2. Repository Data Structure (image omitted)
The repository contains the full HTML of every web page. Each page is compressed using zlib (see RFC1950). The choice of compression technique is a tradeoff between speed and compression ratio. We chose zlib's speed over a significant improvement in compression offered by bzip. The compression rate of bzip was approximately 4 to 1 on the repository as compared to zlib's 3 to 1 compression. In the repository, the documents are stored one after the other and are prefixed by docID, length, and URL as can be seen in Figure 2. The repository requires no

other data structures to be used in order to access it. This helps with data consistency and makes development much easier; we can rebuild all the other data structures from only the repository and a file which lists crawler errors.

4.2.3 Document Index

The document index keeps information about each document. It is a fixed width ISAM (Index sequential access mode) index, ordered by docID. The information stored in each entry includes the current document status, a pointer into the repository, a document checksum, and various statistics. If the document has been crawled; it also contains a pointer into a variable width file called docinfo which contains its URL and title. Otherwise the pointer points into the URLlist which contains just the URL. This design decision was driven by the desire to have a reasonably compact data structure, and the ability to fetch a record in one disk seek during a search.

Additionally, there is a file which is used to convert URLs into docIDs. It is a list of URL checksums with their corresponding docIDs and is sorted by checksum. In order to find the docID of a particular URL, the URL's checksum is computed and a binary search is performed on the checksums file to find its docID. URLs may be converted into docIDs in batch by doing a merge with this file. This is the technique the URLresolver uses to turn URLs into docIDs. This batch mode of update is crucial because otherwise we must perform one seek for every link which assuming one disk would take more than a month for our 322 million link dataset.

4.2.4 Lexicon

The lexicon has several different forms. One important change from earlier systems is that the lexicon can fit in memory for a reasonable price. In the current implementation we can keep the lexicon in memory on a machine with 256 MB of main memory. The current lexicon contains 14 million words (though some rare words were not added to the lexicon). It is implemented in two parts -- a list of the words concatenated together but separated by nulls) and a hash table of pointers. For various functions, the list of words has some auxiliary information which is beyond the scope of this paper to explain fully.

4.2.5 Hit Lists

A hit list corresponds to a list of occurrences of a particular word in a particular document including position, font, and capitalization information. Hit lists account for most of the space used in both the forward and the inverted indices. Because of this, it is important to represent them as efficiently as possible. We considered several alternatives for encoding position, font, and capitalization -- simple encoding (a triple of integers), a compact encoding (a hand optimized allocation of bits), and Huffman coding. In the end we chose a hand optimized compact encoding since it required far less space than the simple encoding and far less bit manipulation than Huffman coding. The details of the hits are shown in Figure 3.

Our compact encoding uses two bytes for every hit. There are two types of hits: fancy hits and plain hits. Fancy hits include hits occurring in a URL, title, anchor text, or meta tag. Plain hits include everything else. A plain hit consists of a capitalization bit, font size, and 12 bits of word position in a document (all positions higher than 4095 are labeled 4096). Font size is represented relative to the rest of the document using three bits (only 7 values are actually used because 111 is the flag that signals a fancy hit). A fancy hit consists of a capitalization bit, the font size set to 7

to indicate it is a fancy hit, 4 bits to encode the type of fancy hit, and 8 bits of position. For anchor hits, the 8 bits of position are split into 4 bits for position in anchor and 4 bits for a hash of the docID the anchor occurs in. This gives us some limited phrase searching as long as there are not that many anchors for a particular word. We expect to update the way that anchor hits are stored to allow for greater resolution in the position and

docIDhash fields. We use font size relative to the rest of the document because when searching, you do not want to rank otherwise identical documents differently just because one of the documents is in a larger font.

Figure 3. Forward and Reverse Indexes and the Lexicon (image omitted)

The length of a hit list is stored before the hits themselves. To save space, the length of the hit list is combined with the wordID in the forward index and the docID in the inverted index. This limits it to 8 and 5 bits respectively (there are some tricks which allow 8 bits to be borrowed from the wordID). If the length is longer than would fit in that many bits, an escape code is used in those bits, and the next two bytes contain the actual length.

4.2.6 Forward Index

The forward index is actually already partially sorted. It is stored in a number of barrels (we used 64). Each barrel holds a range of wordID's. If a document contains words that fall into a particular barrel, the docID is recorded into the barrel, followed by a list of wordID's with hitlists which correspond to those words. This scheme requires slightly more storage because of duplicated docIDs but the difference is very small for a reasonable number of buckets and saves considerable time and coding complexity in the final indexing phase done by the sorter. Furthermore, instead of storing actual wordID's, we store each wordID as a relative difference from the minimum wordID that falls into the barrel the wordID is in. This way, we can use just 24 bits for the wordID's in the unsorted barrels, leaving 8 bits for the hit list length.

4.2.7 Inverted Index

The inverted index consists of the same barrels as the forward index, except that they have been processed by the sorter. For every valid wordID, the lexicon contains a pointer into the barrel that wordID falls into. It points to a doclist of docID's together with their corresponding hit lists. This doclist represents all the occurrences of that word in all documents.

An important issue is in what order the docID's should appear in the doclist. One simple solution is to store them sorted by docID. This allows for quick merging of different doclists for multiple word queries. Another option is to store them sorted by a ranking of the occurrence of the word in each document. This makes answering one word queries trivial and makes it likely that the answers to multiple word queries are near the start. However, merging is much more difficult. Also, this makes development much more difficult in that a change to the ranking function requires a rebuild of the index. We chose a compromise between these options, keeping two sets of inverted barrels -- one set for hit lists which include title or anchor hits and another set for all hit lists. This way, we check the first set of barrels first and if there are not enough matches within those barrels we check the larger ones.

4.3 Crawling the Web

Running a web crawler is a challenging task. There are tricky performance and reliability issues and even more importantly, there are social issues. Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers and various name servers which are all beyond the control of the system.

In order to scale to hundreds of millions of web pages, Google has a fast distributed crawling system. A single URLserver serves lists of URLs to a number of crawlers (we typically ran about 3). Both the URLserver and the crawlers are implemented in Python. Each crawler keeps roughly 300 connections open at once. This is necessary to retrieve web pages at a fast enough pace. At peak speeds, the system can crawl over 100 web pages per second using four crawlers. This amounts to roughly 600K per second of data. A major performance stress is DNS lookup. Each crawler maintains a its own DNS cache so it does not need to do a DNS lookup

before crawling each document. Each of the hundreds of connections can be in a number of different states: looking up DNS, connecting to host, sending request, and receiving response. These factors make the crawler a complex component of the system. It uses asynchronous IO to manage events, and a number of queues to move page fetches from state to state.

It turns out that running a crawler which connects to more than half a million servers, and generates tens of millions of log entries generates a fair amount of email and phone calls. Because of the vast number of people coming on line, there are always those who do not know what a crawler is, because this is the first one they have seen. Almost daily, we receive an email something like, "Wow, you looked at a lot of pages from my web site. How did you like it?" There are also some people who do not know about the robots exclusion protocol, and think their page should be protected from indexing by a statement like, "This page is copyrighted and should not be indexed", which needless to say is difficult for web crawlers to understand. Also, because of the huge amount of data involved, unexpected things will happen. For example, our system tried to crawl an online game. This resulted in lots of garbage messages in the middle of their game! It turns out this was an easy problem to fix. But this problem had not come up until we had downloaded tens of millions of pages. Because of the immense variation in web pages and servers, it is virtually impossible to test a crawler without running it on large part of the Internet. Invariably, there are hundreds of obscure problems which may only occur on one page out of the whole web and cause the crawler to crash, or worse, cause unpredictable or incorrect behavior. Systems which access large parts of the Internet need to be designed to be very robust and carefully tested. Since large complex systems such as crawlers will invariably cause problems, there needs to be significant resources devoted to reading the email and solving these problems as they come up.

4.4 Indexing the Web

Parsing -- Any parser which is designed to run on the entire Web must handle a huge array of possible errors. These range from typos in HTML tags to kilobytes of zeros in the middle of a tag, non-ASCII characters, HTML tags nested hundreds deep, and a great variety of other errors that challenge anyone's imagination to come up with equally creative ones. For maximum speed, instead of using YACC to generate a CFG parser, we use flex to generate a lexical analyzer which we outfit with its own stack. Developing this parser which runs at a reasonable speed and is very robust involved a fair amount of work.

Indexing Documents into Barrels -- After each document is parsed, it is encoded into a number of barrels. Every word is converted into a wordID by using an in-memory hash table -- the lexicon. New additions to the lexicon hash table are logged to a file. Once the words are converted into wordID's, their occurrences in the current document are translated into hit lists and are written into the forward barrels. The main difficulty with parallelization of the indexing phase is that the lexicon needs to be shared. Instead of sharing the lexicon, we took the approach of writing a log of all the extra words that were not in a base lexicon, which we fixed at 14 million words. That way multiple indexers can run in parallel and then the small log file of extra words can be processed by one final indexer.

Sorting -- In order to generate the inverted index, the sorter takes each of the forward barrels and sorts it by wordID to produce an inverted barrel for title and anchor hits and a full text inverted barrel. This process happens one barrel at a time, thus requiring little temporary storage. Also, we parallelize the sorting phase to use as many machines as we have simply by running multiple sorters, which can process different buckets at the same time. Since the barrels don't fit into main memory, the sorter further subdivides them into baskets which do fit into memory based on wordID and docID. Then the sorter, loads each basket into memory, sorts

it and writes its contents into the short inverted barrel and the full inverted barrel.

4.5 Searching

The goal of searching is to provide quality search results efficiently. Many of the large commercial search engines seemed to have made great progress in terms of efficiency. Therefore, we have focused more on quality of search in our research, although we believe our solutions are scalable to commercial volumes with a bit more effort. The google query evaluation process is show in Figure 4.

Figure 4. Google Query Evaluation (image omitted)

To put a limit on response time, once a certain number (currently 40,000) of matching documents are found, the searcher automatically goes to step 8 in Figure 4. This means that it is possible that sub-optimal results would be returned. We are currently investigating other ways to solve this problem. In the past, we sorted the hits according to PageRank, which seemed to improve the situation.

4.5.1 The Ranking System

Google maintains much more information about web documents than typical search engines. Every hitlist includes position, font, and capitalization information. Additionally, we factor in hits from anchor text and the PageRank of the document. Combining all of this information into a rank is difficult. We designed our ranking function so that no particular factor can have too much influence. First, consider the simplest case -- a single word query. In order to rank a document with a single word query, Google looks at that document's hit list for that word. Google considers each hit to be one of several different types (title, anchor, URL, plain text large font, plain text small font, ...), each of which has its own type-weight. The type-weights make up a vector indexed by type. Google counts the number of hits of each type in the hit list. Then every count is converted into a count-weight. Count-weights increase linearly with counts at first but quickly taper off so that more than a certain count will not help. We take the dot product of the vector of count-weights with the vector of type-weights to compute an IR score for the document. Finally, the IR score is combined with PageRank to give a final rank to the document.

For a multi-word search, the situation is more complicated. Now multiple hit lists must be scanned through at once so that hits occurring close together in a document are weighted higher than hits occurring far apart. The hits from the multiple hit lists are matched up so that nearby hits are matched together. For every matched set of hits, a proximity is computed. The proximity is based on how far apart the hits are in the document (or anchor) but is classified into 10 different value "bins" ranging from a phrase match to "not even close". Counts are computed not only for every type of hit but for every type and proximity. Every type and proximity pair has a type-prox-weight. The counts are converted into count-weights and we take the dot product of the count-weights and the type-prox-weights to compute an IR score. All of these numbers and matrices can all be displayed with the search results using a special debug mode. These displays have been very helpful in developing the ranking system.

4.5.2 Feedback

The ranking function has many parameters like the type-weights and the type-prox-weights. Figuring out the right values for these parameters is something of a black art. In order to do this, we have a user feedback mechanism in the search engine. A trusted user may optionally evaluate all of the results that are returned. This feedback is saved. Then when we modify the ranking function, we can see the impact of this change on all previous searches which were ranked. Although far from perfect, this gives us some idea of how a change in the ranking function affects the search results.

5 Results and Performance

Figure 4. Sample Results from Google (image omitted)

The most important measure of a search engine is the quality of its search results. While a complete user evaluation is beyond the scope of this paper, our own experience with Google has shown it to produce better results than the major commercial search engines for most searches. As an example which illustrates the use of PageRank, anchor text, and proximity, Figure 4 shows Google's results for a search on "bill clinton". These results demonstrates some of Google's features. The results are clustered by server. This helps considerably when sifting through result sets. A number of results are from the whitehouse.gov domain which is what one may reasonably expect from such a search. Currently, most major commercial search engines do not return any results from whitehouse.gov, much less the right ones. Notice that there is no title for the first result. This is because it was not crawled. Instead, Google relied on anchor text to determine this was a good answer to the query. Similarly, the fifth result is an email address which, of course, is not crawlable. It is also a result of anchor text.

All of the results are reasonably high quality pages and, at last check, none were broken links. This is largely because they all have high PageRank. The PageRanks are the percentages in red along with bar graphs. Finally, there are no results about a Bill other than Clinton or about a Clinton other than Bill. This is because we place heavy importance on the proximity of word occurrences. Of course a true test of the quality of a search engine would involve an extensive user study or results analysis which we do not have room for here. Instead, we invite the reader to try Google for themselves at http://google.stanford.edu.

5.1 Storage Requirements

Aside from search quality, Google is designed to scale cost effectively to the size of the Web as it grows. One aspect of this is to use storage efficiently. Table 1 has a breakdown of some statistics and storage requirements of Google. Due to compression the total size of the repository is about 53 GB, just over one third of the total data it stores. At current disk prices this makes the repository a relatively cheap source of useful data. More importantly, the total of all the data used by the search engine requires a comparable amount of storage, about 55 GB. Furthermore, most queries can be answered using just the short inverted index. With better encoding and compression of the Document Index, a high quality web search engine may fit onto a 7GB drive of a new PC.

Table 1. Statistics (image omitted)

5.2 System Performance

It is important for a search engine to crawl and index efficiently. This way information can be kept up to date and major changes to the system can be tested relatively quickly. For Google, the major operations are Crawling, Indexing, and Sorting. It is difficult to measure how long crawling took overall because disks filled up, name servers crashed, or any number of other problems which stopped the system. In total it took roughly 9 days to download the 26 million pages (including errors). However, once the system was running smoothly, it ran much faster, downloading the last 11 million pages in just 63 hours, averaging just over 4 million pages per day or 48.5 pages per second. We ran the indexer and the crawler simultaneously. The indexer ran just faster than the crawlers. This is largely because we spent just enough time optimizing the indexer so that it would not be a bottleneck. These optimizations included bulk updates to the document index and placement of critical data structures on the local disk. The indexer runs at roughly 54 pages per second. The sorters can be run completely in parallel; using four machines, the whole process of sorting takes about 24 hours.

5.3 Search Performance

Improving the performance of search was not the major focus of our research up to this point. The current version of Google answers most queries in between 1 and 10 seconds. This time is mostly dominated by disk IO over NFS (since disks are spread over a number of machines).

Furthermore, Google does not have any optimizations such as query caching, subindices on common terms, and other common optimizations. We intend to speed up Google considerably through distribution and hardware, software, and algorithmic improvements. Our target is to be able to handle several hundred queries per second. Table 2 has some sample query times from the current version of Google. They are repeated to show the speedups resulting from cached IO.

Table 2. Search Times (image omitted)

6 Conclusions

Google is designed to be a scalable search engine. The primary goal is to provide high quality search results over a rapidly growing World Wide Web. Google employs a number of techniques to improve search quality including page rank, anchor text, and proximity information. Furthermore, Google is a complete architecture for gathering web pages, indexing them, and performing search queries over them.

6.1 Future Work

A large-scale web search engine is a complex system and much remains to be done. Our immediate goals are to improve search efficiency and to scale to approximately 100 million web pages. Some simple improvements to efficiency include query caching, smart disk allocation, and subindices. Another area which requires much research is updates. We must have smart algorithms to decide what old web pages should be recrawled and what new ones should be crawled. Work toward this goal has been done in (Cho 98). One promising area of research is using proxy caches to build search databases, since they are demand driven. We are planning to add simple features supported by commercial search engines like boolean operators, negation, and stemming. However, other features are just starting to be explored such as relevance feedback and clustering (Google currently supports a simple hostname based clustering). We also plan to support user context (like the user's location), and result summarization. We are also working to extend the use of link structure and link text. Simple experiments indicate PageRank can be personalized by increasing the weight of a user's home page or bookmarks. As for link text, we are experimenting with using text surrounding links in addition to the link text itself. A Web search engine is a very rich environment for research ideas. We have far too many to list here so we do not expect this Future Work section to become much shorter in the near future.

6.2 High Quality Search

The biggest problem facing users of web search engines today is the quality of the results they get back. While the results are often amusing and expand users' horizons, they are often frustrating and consume precious time. For example, the top result for a search for "Bill Clinton" on one of the most popular commercial search engines was the Bill Clinton Joke of the Day: April 14, 1997. Google is designed to provide higher quality search so as the Web continues to grow rapidly, information can be found easily. In order to accomplish this Google makes heavy use of hypertextual information consisting of link structure and link (anchor) text. Google also uses proximity and font information. While evaluation of a search engine is difficult, we have subjectively found that Google returns higher quality search results than current commercial search engines. The analysis of link structure via PageRank allows Google to evaluate the quality of web pages. The use of link text as a description of what the link points to helps the search engine return relevant (and to some degree high quality) results. Finally, the use of proximity information helps increase relevance a great deal for many queries.

6.3 Scalable Architecture

Aside from the quality of search, Google is designed to scale. It must be efficient in both space and time, and constant factors are very important when dealing with the entire Web. In implementing Google, we have seen bottlenecks in CPU, memory access, memory capacity, disk seeks, disk

throughput, disk capacity, and network IO. Google has evolved to overcome a number of these bottlenecks during various operations. Google's major data structures make efficient use of available storage space. Furthermore, the crawling, indexing, and sorting operations are efficient enough to be able to build an index of a substantial portion of the web -- 24 million pages, in less than one week. We expect to be able to build an index of 100 million pages in less than a month.

6.4 A Research Tool

In addition to being a high quality search engine, Google is a research tool. The data Google has collected has already resulted in many other papers submitted to conferences and many more on the way. Recent research such as (Abiteboul 97) has shown a number of limitations to queries about the Web that may be answered without having the Web available locally. This means that Google (or a similar system) is not only a valuable research tool but a necessary one for a wide range of applications. We hope Google will be a resource for searchers and researchers all around the world and will spark the next generation of search engine technology.

7 Acknowledgments

Vitae

(image omitted)

Sergey Brin received his B.S. degree in mathematics and computer science from the University of Maryland at College Park in 1993. Currently, he is a Ph.D. candidate in computer science at Stanford University where he received his M.S. in 1995. He is a recipient of a National Science Foundation Graduate Fellowship. His research interests include search engines, information extraction from unstructured sources, and data mining of large text collections and scientific data.

Lawrence Page was born in East Lansing, Michigan, and received a B.S.E. in Computer Engineering at the University of Michigan Ann Arbor in 1995. He is currently a Ph.D. candidate in Computer Science at Stanford University. Some of his research interests include the link structure of the web, human computer interaction, search engines, scalability of information access interfaces, and personal data mining.

8 Appendix A: Advertising and Mixed Motives

Currently, the predominant business model for commercial search engines is advertising. The goals of the advertising business model do not always correspond to providing quality search to users. For example, in our prototype search engine one of the top results for cellular phone is "The Effect of Cellular Phone Use Upon Driver Attention", a study which explains in great detail the distractions and risk associated with conversing on a cell phone while driving. This search result came up first because of its high importance as judged by the PageRank algorithm, an approximation of citation importance on the web (Page, 98). It is clear that a search engine which was taking money for showing cellular phone ads would have difficulty justifying the page that our system returned to its paying advertisers. For this type of reason and historical experience with other media (Bagdikian 83), we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers.

Since it is very difficult even for experts to evaluate search engines, search engine bias is particularly insidious. A good example was OpenText, which was reported to be selling companies the right to be listed at the top of the search results for particular queries (Marchiori 97). This type of bias is much more insidious than advertising, because it is not clear who "deserves" to be there, and who is willing to pay money to be listed. This business model resulted in an uproar, and OpenText has ceased to be a viable search engine. But less blatant bias are likely to be tolerated by the market. For example, a search engine could add a small factor to search results from "friendly" companies, and subtract a factor from results from competitors. This type of bias is very difficult to detect but could still have a significant effect on the market. Furthermore, advertising income often provides an incentive to provide poor quality search results. For example, we noticed a major search engine would not return a large airline's homepage when the airline's name was given as a query. It so happened that the airline had placed an expensive ad, linked to the query that was its name. A better search engine would not have required this ad, and possibly resulted in the loss of the revenue from the airline to the search engine. In general, it could be argued from the consumer point of view that the better the search engine is, the fewer advertisements will be needed for the consumer to find what they want. This of course erodes the advertising supported business model of the existing search engines. However, there will always be money from advertisers who want a customer to switch products, or have something that is genuinely new. But we believe the issue of advertising causes enough mixed incentives that it is crucial to have a competitive search engine that is transparent and in the academic realm.

9 Appendix B: Scalability

9. 1 Scalability of Google

We have designed Google to be scalable in the near term to a goal of 100 million web pages. We have just received disk and machines to handle roughly that amount. All of the time consuming parts of the system are parallelize and roughly linear time. These include things like the crawlers, indexers, and sorters. We also think that most of the data structures will deal gracefully with the expansion. However, at 100 million web pages we will be very close up against all sorts of operating system limits in the common operating systems (currently we run on both Solaris and Linux). These include things like addressable memory, number of open file descriptors, network sockets and bandwidth, and many others. We believe expanding to a lot more than 100 million pages would greatly increase the complexity of our system.

9.2 Scalability of Centralized Indexing Architectures

As the capabilities of computers increase, it becomes possible to index a very large amount of text for a reasonable cost. Of course, other more bandwidth intensive media such as video is likely to become more pervasive. But, because the cost of production of text is low compared to media like video, text is likely to remain very pervasive. Also, it is likely that soon we will have speech recognition that does a reasonable job converting speech into text, expanding the amount of text available. All of this provides amazing possibilities for centralized indexing. Here is an illustrative example. We assume we want to index everything everyone in the US has written for a year. We assume that there are 250 million people in the US and they write an average of 10k per day. That works out to be about 850 terabytes. Also assume that indexing a terabyte can be done now for a reasonable cost. We also assume that the indexing methods used over the text are linear, or nearly linear in their complexity. Given all these assumptions we can compute how long it would take before we could index our 850 terabytes for a reasonable cost assuming certain growth factors. Moore's Law was defined in 1965 as a doubling every 18 months in processor power. It has held remarkably true, not just for processors, but for other important system parameters such as disk as well. If we assume that Moore's law holds for the future, we need only 10

more doublings, or 15 years to reach our goal of indexing everything everyone in the US has written for a year for a price that a small company could afford. Of course, hardware experts are somewhat concerned Moore's Law may not continue to hold for the next 15 years, but there are certainly a lot of interesting centralized applications even if we only get part of the way to our hypothetical example.

Of course a distributed systems like Gloss (Gravano 94) or Harvest will often be the most efficient and elegant technical solution for indexing, but it seems difficult to convince the world to use these systems because of the high administration costs of setting up large numbers of installations. Of course, it is quite likely that reducing the administration cost drastically is possible. If that happens, and everyone starts running a distributed indexing system, searching would certainly improve drastically.

Because humans can only type or speak a finite amount, and as computers continue improving, text indexing will scale even better than it does now. Of course there could be an infinite amount of machine generated content, but just indexing huge amounts of human generated content seems tremendously useful. So we are optimistic that our centralized web search engine architecture will improve in its ability to cover the pertinent text information over time and that there is a bright future for search.

Other references

References:

Best of the Web 1994 -- Navigators
http://botw.org/1994/awards/navigators.html
Bill Clinton Joke of the Day: April 14, 1997.
http://www.io.com/~cjburke/clinton/970414.html.
Bzip2 Homepage http://www.muraroa.demon.co.uk/
Google Search Engine http://google.stanford.edu/
Harvest http://harvest.transarc.com/
Mauldin, Michael L. Lycos Design Choices in an Internet Search Service, IEEE **Expert** Interview
http://www.computer.org/pubs/expert/1997/trends/x1008/mauldin.htm
The Effect of Cellular Phone Use Upon Driver Attention
http://www.webfirst.com/aaa/text/cell/cell0toc.htm
Search Engine Watch http://www.searchenginewatch.com/
RFC 1950 (zlib) ftp://ftp.uu.net/graphics/png/documents/zlib/zdoc-index.html
Robots Exclusion Protocol:
http://info.webcrawler.com/mak/projects/robots/exclusion.htm
Web Growth Summary: http://www.mit.edu/people/mkgray/net/web-growth-summary.html
Yahoo! http://www.yahoo.com/
(Abiteboul 97) Serge Abiteboul and Victor Vianu, Queries and Computation on the Web. Proceedings of the International Conference on Database Theory. Delphi, Greece 1997.
(Bagdikian 97) Ben H. Bagdikian. The Media Monopoly. 5th Edition. Publisher: Beacon, ISBN: 0807061557
(Chakrabarti 98) S.Chakrabarti, B.Dom, D.Gibson, J.Kleinberg, P. Raghavan and S. Rajagopalan. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. Seventh International Web Conference (WWW 98). Brisbane, Australia, April 14-18, 1998.
(Cho 98) Junghoo Cho, Hector Garcia-Molina, Lawrence Page. Efficient Crawling Through URL Ordering. Seventh International Web Conference (WWW 98). Brisbane, Australia, April 14-18, 1998.
(Gravano 94) Luis Gravano, Hector Garcia-Molina, and A. Tomasic. The Effectiveness of GIOSS for the Text-Database Discovery Problem. Proc. of the 1994 ACM SIGMOD International Conference On Management Of Data, 1994.
(Kleinberg 98) Jon Kleinberg, Authoritative Sources in a Hyperlinked Environment, Proc. ACM-SIAM Symposium on Discrete Algorithms,

1998.

(Marchiori 97) Massimo Marchiori. The Quest for Correct Information on the Web: Hyper Search Engines. The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.

(McBryan 94) Oliver A. McBryan. GENVL and WWWW: Tools for Taming the Web. First International Conference on the World Wide Web. CERN, Geneva (Switzerland), May 25-26-27 1994.
http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps

(Page 98) Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Manuscript in progress.
http://google.stanford.edu/~backrub/pageranksub.ps

(Pinkerton 94) Brian Pinkerton, Finding What People Want: Experiences with the WebCrawler. The Second International WWW Conference Chicago, USA, October 17-20, 1994.
http://info.webcrawler.com/bp/WWW94.html

(Spertus 97) Ellen Spertus. ParaSite: Mining Structural Information on the Web. The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.

(TREC 96) Proceedings of the fifth Text REtrieval Conference (TREC-5). Gaithersburg, Maryland, November 20-22, 1996. Publisher: Department of Commerce, National Institute of Standards and Technology. Editors: D. K. Harman and E. M. Voorhees. Full text at:
http://trec.nist.gov/

(Witten 94) Ian H Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. New York: Van Nostrand Reinhold, 1994.

(Weiss 96) Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. Proceedings of the 7th ACM Conference on Hypertext. New York, 1996.

Examiner submission date
20020429.

Image(s)

[PDF Adobe icon]

[save] locally as: |PDF document ▼| search strategy: |do not include the search strategy ▼|
[previous documents] [next documents] [order]

00620511 20011113317B3428 (USE FORMAT 7 FOR FULLTEXT)
**RuleSpace Wins PC Magazine Technical Excellence Award-Contexion(TM)
Services Selected as Most Innovative Internet Software and Service Solution**
Business Wire
Tuesday, November 13, 2001 13:47 EST
JOURNAL CODE: BW LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 734

...Services scales successfully with
the exponential growth of the Web by employing techniques based on **neural
network** technology that categorize Web content continually at record
speed.
Deployed by leading companies such as...

...number of sites increasing daily, it's a tough job. RuleSpace
Contexion Services uses patented **neural network** technology to work
24/7,
categorizing content . . . Earlier this year, AOL incorporated Contexion
Services into...
...said Dan Lulich, CTO at
RuleSpace and a key developer of the world's first **neural network**
computer.
"We are honored to be recognized as a provider of revolutionary technology
that is...

...categorizing Web
sites - analyzing pages, site structure, subdirectories and links on the
site.
The unique **features** --patterns of letters, words and phrases--of a given
Web
page are analyzed in real-time by the application of patented **neural
network**
technology and pattern recognition **techniques** to determine if the page
belongs
to specific content categories. The results of the analysis are then
aggregated to infer an overall category **rating** for **Web sites** and
subdirectories and stored in the largest **database** of pre-categorized
inappropriate Web sites and subdirectories available today.

RuleSpace provides enterprise access control...

 10/3,K/7   (Item 4 from file: 610)

00551313 20010710191B1714 (USE FORMAT 7 FOR FULLTEXT)
**Cerberian Adopts RuleSpace Technology to Enable Next-Generation Internet
Access Management-Cerberian Internet Manager Offers Automated Web Content
Recognition in Its Entirely Web-Based Service**
Business Wire
Tuesday, July 10, 2001 08:04 EDT
JOURNAL CODE: BW LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE

WORD COUNT: 832

...a given
Web page are analyzed in real-time by the application of patent-pending
neural
 network technology and pattern recognition **techniques** to determine if
the page
belongs to specific content categories.

This approach enables Contexion Services...

...and subdirectories available today.
The product continually analyzes millions of Web sites through an automated
 process by considering their page content, page structure, site
relationship
and links to other known sites.

The results of the analysis are then aggregated to infer an overall
category
 rating for **Web   sites** and subdirectories, then stored in the **database**
. The
combination of Contexion Services' real-time categorization and automated
retrieval of site category ratings from Contexion Services' unmatched
database
of pre-categorized content results in the highest possible degree of
filtering
precision.

    About Cerberian...

    10/3,K/8   (Item 5 from file: 610)
DIALOG(R)File 610:Business Wire
(c) 2005 Business Wire. All rts. reserv.

00273712 20000508129B4263   (USE FORMAT 7 FOR FULLTEXT)
RuleSpace Categorizing the World Wide Web; Over Two Million Web Sites
Categorized Using Instant Web Content Recognition Technology
Business Wire
Monday, May 8, 2000 06:19 EDT
JOURNAL CODE: BW LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 1,017

...information through
        facilitating vertical search and content discovery.

    The Contexion Approach

Contexion uses two complementary **techniques** for categorizing the Web:
content
and context analysis. The content analysis component of Contexion examines
all
textual **attributes** of an incoming Web page in real-time including words,
phrases, meta data and page structure. Once the **attributes** have been
recognized, they are compared to a "rulespace" -- a unique set of defining
 attributes for each category created by RuleSpace's **neural   networks** .
If the

attributes of the Web page resemble those of the rulespace, the page is classified as a...

...data in a fraction of the time it would take for manual
review.
The second **technique** , context analysis, infers the category **rating** of a
Web
page through association to its subdirectory and parent Web site. Web
pages
themselves often do not...

...words. To address this challenge, Contexion continually categorizes the
Web
offline to determine the category **ratings** of all known Web **sites** and
high-level subdirectories. These **ratings** provide context for any **Web**
page . The
Category Name Server, which stores these ratings, currently contains
millions
of pre-categorized Web sites -- effectively representing the entire Web.
The combination of both **techniques** -content and context analysis-results
in
instant Web content recognition, providing scalable, precise categorization
of...

  10/3,K/9   (Item 1 from file: 613)
DIALOG(R)File 613:PR Newswire
(c) 2005 PR Newswire Association Inc. All rts. reserv.

00673523 20011109DEF008 (USE FORMAT 7 FOR FULLTEXT)
BudgetLife Launches Expert **Quoting** System
PR Newswire
Friday, November 9, 2001 07:57 EST
JOURNAL CODE: PR LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
DOCUMENT TYPE: NEWSWIRE
WORD COUNT: 427

BudgetLife Launches Expert **Quoting** System

TEXT:
Interlinx, LLC today announced the
BudgetLife **Expert** Quoting **System** , and implemented the new software at
its Web
site at http://www.budgetlife.com . The **Expert** Quoting **System** uses
proprietary technology to help consumers locate life insurance **rates** for
which
they qualify.
   "Most **Web** **sites** selling life insurance today simply list the lowest
rates
being offered by a group of...
...their weight or family health history. We solved
this problem by programming the companies' underwriting **rules** right into
the
software."
   According to Mr. Burt, consumers using the old **method** would often
apply
with the company showing the lowest rate, only to find out eight...